STATISTICAL PROBLEMS IN MEASURING SURFACE OZONE AND MODELLING ITS PATTERNS

Paul Stewart Hutchison

A Thesis Submitted for the Degree of MPhil
at the
University of St Andrews

1996

Statistical Problems In Measuring Surface Ozone And Modelling Its Patterns

By

Paul Stewart Hutchison

Being a Thesis submitted to the
University of St Andrews in candidature for
the degree of Doctor of Philosophy

September 1995

Copyright Declaration

A UNRESTRICTED

DECLARATION

a) I, Paul Stewart Hutchison, hereby certify that this thesis has been composed by myself, that it is a record of my own work, and it has not been accepted in partial of complete fulfilment of any other degree or professional qualification.

Signed_ ____ Date 5 OCT 1995

b) I was admitted to the Faculty of Science of the University of St Andrews, under Ordinance General No 12 on October 1991, and as a candidate for the degree of Ph.D. on October 1992.

Signed Date 5 OCT 1995

c) I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate to the degree of Ph.D.

Signature of Supervisor_ Date 6 Oct 1995

Abstract

The Thesis examines ground level air pollution data supplied by ITE Bush, Penicuik, Midlothian, Scotland. There is a brief examination of sulphur dioxide concentration data, but the Thesis is primarily concerned with ozone. The diurnal behaviour of ozone is the major topic, and a new methodology of classification of 'ozone days' is introduced and discussed.

In chapter 2, the inverse Gaussian distribution is considered and rejected as a possible alternative to the standard approach of using the lognormal as a model for the frequency distribution of observed sulphur dioxide concentrations.

In chapter 3, the behaviour of digital gas pollution analysers is investigated by making use of data obtained from two such machines operating side by side. A time series model of the differences between the readings obtained from the two machines is considered, and possible effects on modelling discussed.

In chapter 4, the changes in the diurnal behaviour of ozone over a year are examined. A new approach involving a distortion of the time axis is shown to give diurnal ozone curves more homogeneous properties and have beneficial effects for modelling purposes.

Chapter 5 extends the analysis of the diurnal behaviour of ozone begun in chapter 4 by considering individual 'ozone days' and attempting to classify them as one of several typical 'types' of day. The time distortion method introduced in chapter 4 is used, and a new classification methodology is introduced for considering data of this type. The statistical properties of this method are discussed in chapter 6.

Acknowledgements

Thanks are especially due to Professor Richard Cormack at the Department of Statistics, St Andrews University, St Andrews and Ron Smith at the Institute of Terrestrial Ecology, Bush Estates, Penicuik, Midlothian for all their help and time that went into the work described in this thesis.

I would also like to thank Professor David Fowler at the Institute of Terrestrial Ecology at Bush for his help with gathering of data and other information and interpretation of the results.

In addition, Professor Steve Buckland at St Andrews provided invaluable help with the final stages of submission and printing.

Table of Contents

Table of Contents (continued)

Table of Contents (continued)

Table of Contents (continued)

Chapter 1: Thesis Introduction

With regular reports in the media, public interest has been increasing in the dangers and results of pollution and other environmental damage. During the last 40 years, both in Europe and North America, numerous examples of visible damage to vegetation due to ground level air pollutants have been observed, especially near large industrial sources. These factors have led to an increasing amount of research into the causes and effects of man-made air pollution. Public opinion calls for legislation aimed at preventing the observed problems, but in order to legislate effectively it is essential to first understand the mechanisms that govern air pollutant levels and the resultant damage to the environment, people and materials. Much activity has been focused on the harmful effects of various air pollutants on plants, especially agricultural crops, where the resultant economic damage has the potential of making a large impact. Research on the effects of individual air pollutant gases has identified sulphur dioxide (SO_2) and nitrogen oxides (NOx) as those responsible for much of the observed plant damage. Ozone (O_3) has also been identified as an important pollutant gas.

These gases can be divided into two types: primary and secondary pollutants. Primary pollutants (SO_2, NOx) are emitted directly into the atmosphere, and as a result the observed levels of man-made pollution can be linked directly to the levels of the man-made emissions. Secondary pollutants (O_3) differ from primary pollutants in that there are little or no direct man-made emissions. They are generated by a series of photochemical reactions that occur in air already polluted with primary pollutant gasses termed precursors. Understanding the observed level is thus complicated considerably, as it will be the result of a number of different processes, the rates of which can be affected by many different factors including location, weather, time of day and the levels of various precursors. Another factor which must be considered if the effect of man-made emissions is to be assessed properly is that the pollutant gasses will have a 'natural' background level that would exist if there were no man-made pollution present. However, estimation of this background level is not a trivial task as it is masked by all man-made emissions.

1.1: The Atmosphere

As this thesis is concerned with air pollution, a brief description of the atmosphere is warranted. About 85% of the total atmospheric mass and practically all the water vapour exist in the troposphere which extends from the planet's surface to about 8 kms above ground level at the poles and 15 kms at the equator. The band of air at lowest altitudes, up to about 1 km, is termed the boundary layer. With reference to ozone the stratosphere, which exists above the troposphere to a height of about 50 kms above ground level is also important, as it is within this layer that most ozone formation occurs.

1.2 : Recording Methods

In the UK, the levels of a large number of different ground level pollutant gases, including O_3, SO_2 and NO_x are continuously monitored and recorded by networks of sites covering most of the country. Data from these sites are recorded in several different ways. Sometimes a daily, hourly or 15 minute average reading will be recorded, in other cases the observations may be maxima over similar time periods. These types of data are required for separate analyses of average or maximum behaviour. It is also common to record 'spot values' at regularly spaced time intervals. Most of the data used in this thesis have been recorded as 15 minute averages, which can readily be aggregated to form hourly or daily averages. Measurements are given in parts per billion (ppb), where 1ppb represents a volume mixing ratio of 1 volume of pollutant in 10^9 volumes of air.

Ozone Monitoring In The UK

Analysers which continuously record ambient ozone levels have only been available since about 1970. Prior to this, measurements were obtained by using the ability of ozone to oxidise potassium iodide to iodine. However, as other strong oxidising agents which can also be found in the atmosphere can also cause this reaction, these readings are referred to as measurements of 'total oxidant' and are not ozone specific. In addition, the presence of SO_2 in the sampled air stream will reduce the measured ozone level. As a result of these factors, measurements of ozone levels prior to 1970 are unreliable.

Continuous measurements of ozone levels have generally been made in the UK using analysers which work on one of two distinct principles: The first method (ethylene chemiluminescence) detects photons produced by the reaction which occurs between ethylene and ambient ozone when ethylene is mixed with ambient air in a reaction chamber. This technique is ozone specific, has rapid response and good precision, but instruments based on this principle must be calibrated against a known ozone concentration. This calibration has suffered some problems in the past owing to the problems involved in obtaining a standard ozone source. However these have been overcome since about the mid 1980's. The second method (ultraviolet absorption) utilises the intense ultraviolet absorption band of ozone. Ambient air and then reference ozone-free air are passed sequentially between an ultraviolet light source and a detector. The ratio of the different light intensities transmitted through the two types of air can be used to calculate the ozone concentration using the Beer-Lambert absorption law. The measurement is sensitive to temperature and pressure but sensors can be included to allow the analyser to compensate automatically. This technique is also ozone specific and has about the same precision as ethylene-chemiluminescence but requires no supply of ethylene gas and gives better calibration specifications.

There has been a network of ozone analysing machines operating in the UK since about 1971. A database exists extending back to 1972. As of 1993, data are held for a total of 46 sites located across the country, both in rural and urban areas. These sites are catalogued in both PORG 1993 and 1987. Much of the data used in this thesis are from ITE Bush, site number 5 in the list given in both PORG references. The bulk of the data analysed in this thesis are measurements on levels of ground level ozone, measured in parts per billion (ppb), at ITE, Bush Estate, Midlothian. Data are available in the form of 15 minute average ozone levels, from one or both of two analysing machines, both of which measure ozone by the UV absorption method. The fact that the observations are sometimes available from two machines operating side by side allows an investigation of the differences between the measurements obtained from them, and any implications for data analysis. This investigation is carried out in chapter 3.

1.3 : Primary Pollutants : SO_2 , NO_x

The distribution of observed ground level concentrations is dominated by man-made emissions, due mainly to industrial (and domestic) sources some distance upwind. Nitrogen oxides and sulphur dioxide are produced primarily from high temperature combustion of fossil fuels. In the UK during 1991, vehicles accounted for 51% of NO_x emissions and power stations accounted for 26%. The remaining 23% of emissions were divided between other transport and industrial sources, with only 3% being attributed to domestic sources (PORG 1993).

SO_2- cyclic variations in observed concentrations

Observed concentrations of SO_2 exhibit strong cyclic variation, on both a diurnal and annual level. These cycles are a result of various factors. The annual cycle for primary pollutants reflects the demand for energy throughout the year, whereas the diurnal cycle depends more on meteorological and chemical factors (Fowler & Cape 1982). The cycles can be very site dependent, as a result of whatever sources are nearby. For example if examining the diurnal cycle from a site near to a road, peaks can often be observed at 'rush hour' times.

Annual cycle

For Europe, the cycle typically exhibits a maximum during the winter, reflecting the change in demand for energy (Fowler & Cape 1982). Figure 1.1 gives the annual cycle of SO_2 levels observed at ITE Bush in 1990.

Diurnal cycle

The diurnal cycle of observed SO_2 and NO_x levels typically follow a very similar pattern to each other. The diurnal cycle is generated by a combination of factors, including rates of emissions, deposition rates and meteorological conditions (Fowler & Cape 1982). Figure 1.2 gives the average diurnal cycle of SO_2 from ITE Bush in 1990.

Fig 1.1 : SO2 Annual Cycle 1990 Bush

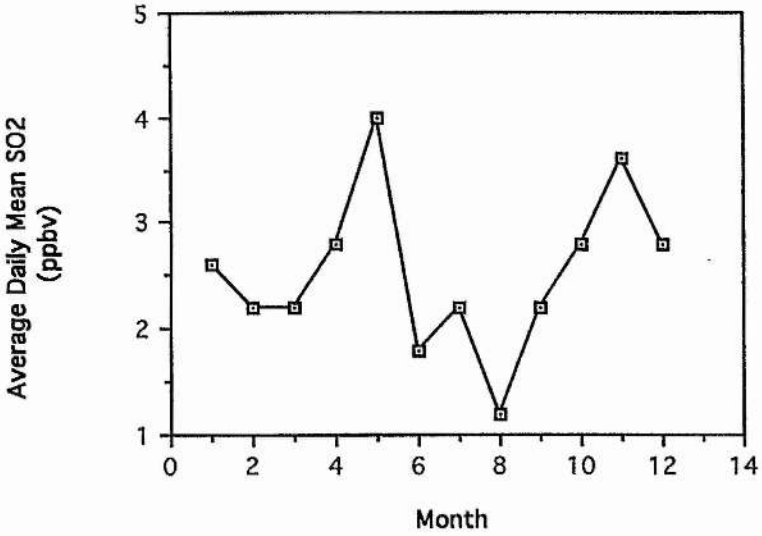
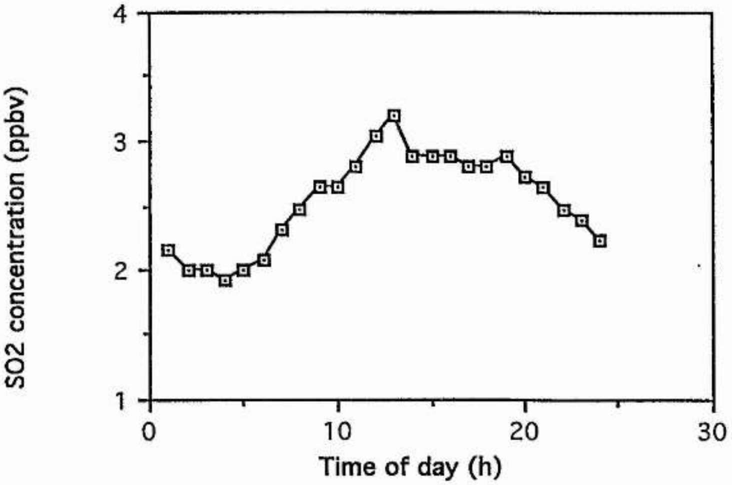


Fig 1.2 : Average SO2 Diurnal Cycle
1990 Bush



SO_2-Frequency distribution

One method for summarising the behaviour of pollutant levels has been to ignore the time ordering of the observations and to fit a frequency distribution to the set of all measurements of pollutant levels obtained over a certain period, often a year. This distribution summarises the behaviour of the pollutant over time, and can be used to estimate the frequency of occurrences of different pollutant levels. These frequencies can be used in conjunction with models for damage to plants, people or materials as a function of pollutant levels to estimate current levels of damage. Owing to the fact that the time structure of the data has been ignored, this approach at best merely provides a statistical description of pollutant levels and gives no understanding of the processes governing these. Despite this, such descriptions of the data can still prove useful, and often form the basis for air quality control legislation, particularly in North America (Smith, Fowler & Cape 1989). The use of a frequency distribution as a summary of pollutant concentrations is further complicated by the fact that data are obtained from a large network of different sites giving different types of observations. Observations may be given in the form of spot values at regularly spaced intervals, averages or maxima are also often recorded for every 15 minutes, hour or whole day. The form of the data will affect the model used, for example methods constructed with 15 minute averages may not work with data in the form of hourly or daily averages. Maxima are best described by the Gumbel extreme value distribution (Smith, Fowler & Cape 1989), and will not be considered in this thesis. Data used in this thesis are in the form of 15 minute average SO_2 levels from ITE Bush over 1990.

The frequency distribution of observed SO_2 concentrations (assumed to be averages over given time periods) can often be well described in the central portion, say between the 20th and 95th percentiles, by a lognormal distribution (Smith Fowler & Cape 1989). Much analysis follows the work of Larsen (1973, 1974), which is based on an empirical model for the data based on the following three properties:

- Frequency distribution tends to be lognormal
- Median averaged concentration proportional to a power of length of averaging interval.
- Maximum averaged concentration proportional to an inverse power of length of averaging interval.

This method allows for prediction of median and maximum values for a given averaging time if the data are available for another averaging time. For example, if daily mean values are available over a sufficiently long period of time (usually several years), then the hourly median and maximum and the annual median and maximum can be predicted. The design of air quality control legislation, especially in North America is frequently based on this approach, as they are often expressed in terms of average concentrations which should not be exceeded for more than one hour or one day per year.

Larsen's method depends on the assumption that the lognormal distribution can be used to describe the frequency distribution of observed pollutant gas levels. Smith, Fowler & Cape (1989) investigate this, and demonstrate that the lognormal underestimates the frequency of the high SO_2 concentrations. This could have serious consequences for investigations into damage due to SO_2 pollution, as it is these high concentrations that will be potentially most damaging. The Inverse Gaussian distribution will be considered as a possible alternative to the lognormal in chapter 2 of this thesis, as it has a heavier upper tail than the lognormal distribution.

1.4 : Secondary Pollutants: O_3

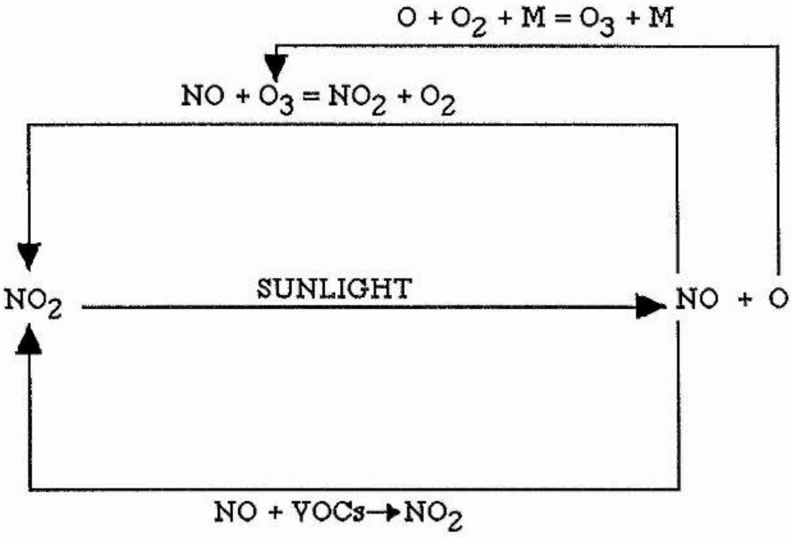
Ozone is produced in polluted air by a complex series of photochemical reactions. The rate of these is largely dependent on weather conditions and the levels of some other (primary) pollutants. There are also large reservoirs of ozone in the stratosphere which in certain weather conditions can be transported into the troposphere, leading to an increase in ground level ozone (PORG 1987).

Photochemical Ozone Production

Ozone is formed in the troposphere in a complex photochemical reaction with the by-products of fossil fuel combustion. A simplified form of this reaction is illustrated in figure 1.3. Sunlight converts NO_2 to NO. The resulting oxygen atom combines with O_2 to make O_3. O_3 also reacts with NO to form NO_2, causing a 'photostationary state' to exist among NO_2, O_3 and NO. When volatile organic compounds (VOCs) and certain other molecules are present, they promote reactions with NO, blocking the reaction with O_3. This causes a build-up of ozone. Sources of VOCs include car exhausts, evaporation of petrol and other solvents, chemical manufacturing and petroleum refining. In some areas, natural production of VOCs from vegetation and animals can also be a major source. (PORG 1993)

Figure 1.3 : Ozone photochemical reaction

M: Any molecule capable of dissipating the released energy eg. Nitrogen, Oxygen

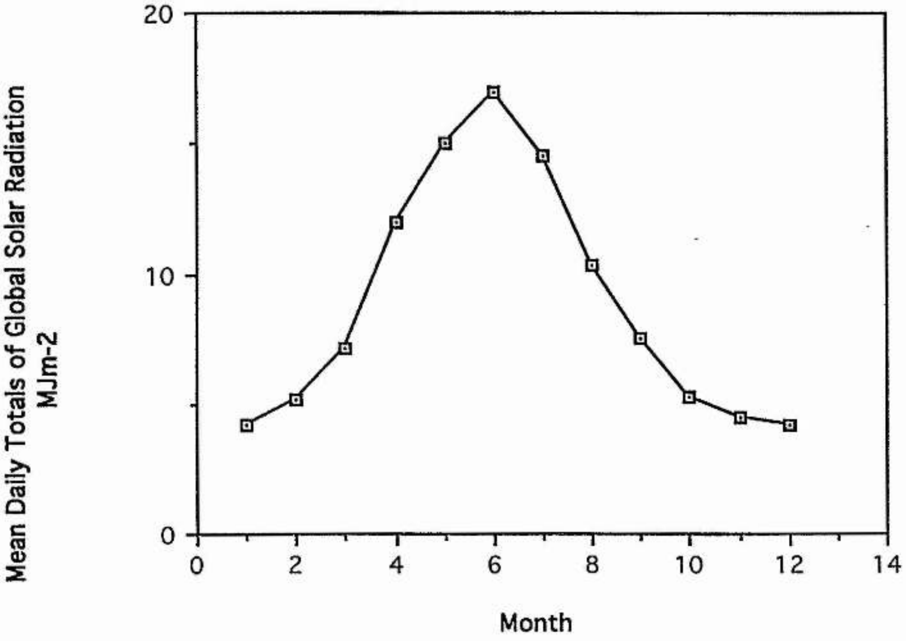


The amount of ozone resulting from this reaction is thus dependent on the levels of NO2 and VOCs present. As the reaction is driven by sunlight, the ozone level will also be weather dependent. There are four main requirements on the weather for ozone formation (PORG 1993):

- Sunshine to drive the chemical reactions.
- Low wind speeds to inhibit dispersion.
- Restricted boundary layer depths, to allow the build up of precursors (eg. NO_2, NOx)
- Air temperature above 20°C to enhance evaporation of hydrocarbon precursors and promote certain chemical reactions.

Figure 1.4 shows a typical graph of mean daily totals of solar radiation against time of year for a site in the UK (PORG 1987). The pronounced annual cycle, differing by a factor of three between January and June, will have a marked effect on the frequency of

Fig 1.4 : Typical Behaviour of Solar Radiation over a year in the UK



photochemical production over a year. As there is significantly more solar radiation in the summer, so too there will be much more photochemical ozone production.

Ozone Depletion

Ozone is deposited readily onto vegetation and soil, in a process termed dry deposition. The mechanism of deposition can be divided into two stages. Turbulent diffusion transports ozone to within a millimetre or two of the absorbing surface and then surface processes determine the mechanism of uptake. Rates of turbulent diffusion are determined primarily by wind velocity and the aerodynamic roughness of the surface. For example the transport rate above a rough surface such as a forest will be greater than that over a smooth surface such as short grass. Another factor which affects the rate of turbulent diffusion is the vertical profile of air temperature; unstable conditions lead to larger rates of turbulent diffusion and vice versa. For vegetation, the most important terrestrial surface, uptake is primarily within the leaves following stomatal absorption. As stomatal opening follows a marked diurnal cycle, with stomata opening in the morning and closing in the evening, so too does the deposition rate. This cycle does however often become obscured in the field owing to increases in atmospheric resistances at night.

Stratospheric Ozone

All atmospheric ozone is produced photochemically. Most is formed in the stratosphere, giving rise to what is termed the ozone layer. Usually there is little exchange between the stratosphere and the troposphere. However given the correct atmospheric conditions, intrusions of stratospheric air, rich in ozone, can enter the middle and lower troposphere as far as the atmospheric boundary layer. The weather conditions which lead to such incursions occur mainly in springtime in the UK. (PORG 1993)

Annual Cycle

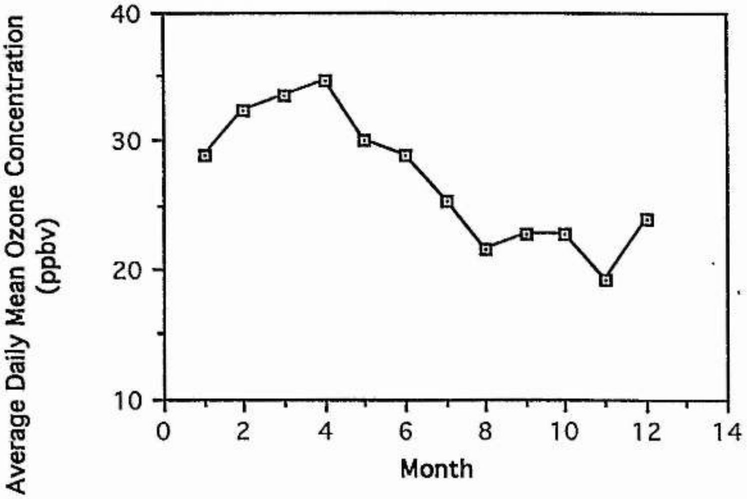
Due to the weather dependence of the processes affecting ozone production and depletion, with highest production occurring during the summer, ozone levels exhibit strong cyclic variations on a seasonal level. Typically, concentrations will peak during the spring, as a result of a similar peak in stratosphere / troposphere exchange and an increase in the frequency of days where photochemical production occurs. There is also typically a drop in levels during August / September as the frequency of days in which photochemical production occurs is reduced.

The annual cycle of ozone levels for ITE Bush in 1990 is illustrated in figure 1.5. This follows the typical pattern described above.

Diurnal cycle

Ozone levels also exhibit strong variation on a diurnal basis. This is due both to the meteorological dependence of the processes governing ozone production, and the diurnal cycle observed in many of the primary pollutant precursors. The average diurnal cycle over a year typically exhibits a 'hump' starting about 8 am and finishing about 7 pm. The diurnal cycle is a common feature of analyses of ozone monitoring data and has been shown to vary from site to site and according to the time of year (PORG 1987). The exact nature of the changes in the diurnal cycle from site to site and over the year has not been examined closely. These changes form one of the major topics of this thesis. In chapter 4, the differences between the average diurnal cycle for each month of a year are investigated. In chapter 5 the behaviour of ozone over individual days is examined, and a new classification methodology is introduced which seeks to identify several typical types of ozone day. The statistical properties of this method are discussed in chapter 6.

Fig 1.5 : Ozone Annual Cycle 1990 Bush



1.5 : Effects Of Air Pollutants

Effects Of Ozone

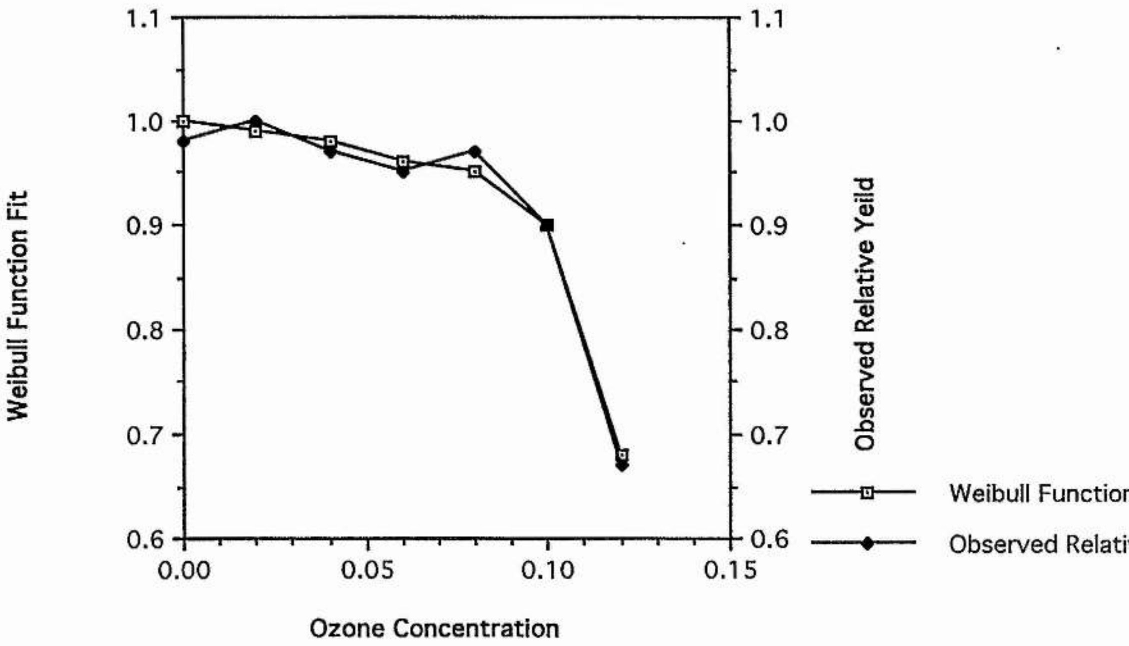
Ozone is a toxic gas which has deleterious effects upon human health, vegetation and materials. If inhaled by humans in sufficient quantities, especially when taking heavy exercise, damage can be caused to the lungs and respiratory tract. Numerous examples of damage to vegetation have been demonstrated, although the mechanisms which govern plant damage have been found to be susceptible to a large number of different factors making categorical statements about the ways in which ozone affects plants difficult. Ozone has also been demonstrated to cause damage to materials. Estimates suggest that the annual damage by ozone in the USA to elastomers, textiles and paints in the 1970s was about $2500 million. Ozone also affects climate as it is a greenhouse gas. Increases in its concentration in the troposphere will lead to warming of the surface-troposphere system. (PORG 1993)

Modelling Pollutant Damage

The most comprehensive study carried out to date has been in North America by NCLAN (the National Crop Loss Assessment Network) which was begun in 1980 to develop an understanding of the effects of pollutants (primarily ozone) on agricultural crops. They have studied crops growing in open topped chambers, exposed to controlled amounts of pollutants. Some work has been carried out by growing crops in ambient levels of pollutants and comparing the yield obtained against that of crops growing in chambers with the pollutants removed. This approach is useful only for assessing damage due to current levels of pollution, but a study using widely varying levels of pollutants will be more useful for modelling purposes. This approach allows the plant damage to be linked by some function to the pollutant dose and allows assessment of possible future damage under many different scenarios and calculation of economic and environmental gains for given reductions in pollutant levels. However, such experiments are difficult to construct well, and ideally are conducted over a long period of time. In addition the term 'pollutant dose' has to be carefully defined. Many possibilities exist, for example daily mean concentrations or time spent above a threshold level. The best modelling approach depends on the mechanisms which govern damage in a particular situation, however there are many gaps in knowledge of these.

The dose response function chosen by NCLAN is a form of the Weibull function, which is used in reliability studies to model the failure of systems. An example of the relationship between yield and ozone levels is illustrated in figure 1.6. The observed means and Weibull function fit are shown. One problem with this is that the Weibull function is a decreasing one, and as such cannot show the increase in yield from zero to ambient levels that are sometimes observed. (Rawlings, Lesser & Dassel). It had been planned when this thesis was begun to spend some time examining ways of improving this approach, either by modifying the Weibull function or by finding some new model. Given more time this would have been carried out.

Fig 1.6 : Typical Dose Response Curve



Chapter 2: SO_2 : Frequency distribution

2.1: Introduction

The usual analysis of data from monitoring of pollutant gases assumes that the frequency distribution of observed concentrations can be described by a lognormal distribution (Smith, Fowler & Cape 1989), (Fowler & Cape 1982).

However, the fit of the lognormal distribution in the upper tail of the distribution is poor, with the lognormal under-predicting the frequency of higher concentrations (Smith, Fowler & Cape 1989). As damage occurs mainly at these higher concentrations, this under-prediction gives rise to concern over the use of the lognormal based approach to legislate air quality standards.

The inverse Gaussian distribution has been suggested as a possible alternative to the lognormal distribution for the frequency distribution of SO_2 levels. The upper tail of the inverse Gaussian is heavier than that of the lognormal and may be able to improve this lack of fit to the data.

This chapter seeks to investigate the possibility of using the inverse Gaussian as a substitute for the lognormal by fitting both to a set of SO_2 concentrations and comparing the results, with particular attention to the fit in the upper tail.

2.2: Methods

The Inverse Gaussian distribution has the probability density function (Chhikara & Folks 1989):

$$f(x;\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi}}\ x^{-3/2} \exp\left(\frac{-\lambda(x-\mu)^2}{2\mu^2 x}\right) \qquad x>0 \qquad \textbf{(2.1)}$$

The maximum likelihood estimators mle(μ) and mle(λ) of μ and λ for a random sample X_1, X_2, ..., X_n from an inverse Gaussian distribution (Chhikara & Folks 1989) are given by:

$$\text{mle}(\mu) = \mathbf{X} \qquad \textbf{(2.2)}$$

$$\frac{1}{\text{mle}(\lambda)} = \frac{1}{n} \sum_{1}^{n} \left(\frac{1}{X_i} - \frac{1}{\mathbf{X}}\right) \qquad \textbf{(2.3)}$$

$$\text{Where } \mathbf{X} = \frac{1}{n} \sum_{1}^{n} X_i$$

Both the inverse Gaussian and 2 parameter lognormal distributions have been fitted by maximum likelihood to a data set consisting of 15 minute averages of SO_2 concentrations (ppbv) at ground level at ITE Bush over 1990. This data set contains 33371 observations, with 1667 missing observations spread randomly throughout the year.

Two standard tests were used to assess the fit of both distributions: the Kolmogorov-Smirnov test, and the Cramer-VonMises test.

A modified form of the Anderson-Darling test introduced in Sinclair, Spurr & Ahmad (1990), which gives more weight to deviations in the upper tail of the distribution has also been used. This test is appropriate given the importance of the fit of the distribution to the upper tail in this case.

This test statistic is calculated by the formula:

$$AU_n^2 = \frac{n}{2} - 2\sum_{j=1}^{n} F(x(j)) - \sum_{j=1}^{n} \left[2 - \frac{(2j - 1)}{n}\right] \log[1 - F(x(j))] \qquad \textbf{(2.4)}$$

Where $F(x_{(j)})$ is the cumulative distribution function of the distribution fitted, evaluated at the j^{th} ordered observation. The critical value at significance level 0.05 as $n \to \infty$ is 0.432.

2.3: Results

The parameters of the inverse Gaussian distribution were estimated by maximum likelihood. The estimates were μ=2.5440 and λ=1.3118.

The 2 parameter lognormal distribution was also fitted. The parameters were estimated as μ=0.4890 and σ=0.96753.

The values of the three test statistics were calculated for both distributions. The results are summarised in table 2.1:

Table 2.1: Values of lack of fit test statistics

Test	Lognormal	Inverse Gaussian	Critical Value
Kolmogorov-Smirnov: D	0.043	0.133	0.007
Cramer-VonMises: T	18.89	199.04	0.461
Mod Anderson-Darling: AU_n^2	41.27	351.37	0.432

Table 2.1 indicates that the inverse Gaussian does not fit the data as well as the lognormal. In addition, the values of AU_n^2 indicate that the lognormal fits the upper tail considerably more accurately. The results confirm that the lognormal does not fit the data particularly well, as the values of all three test statistics are greater than their critical values.

The behaviour of the distributions in the upper tail of the distribution has been mentioned to be of interest. To investigate this further, table 2.2 summarises the behaviour of observations greater than 5 ppb (about 11% of the data).

Table 2.2: Number of observations recorded and predicted above 5 ppb

Obs. Values	Number Observed	Number Predicted	
(ppb)		Inverse Gaussian	Lognormal
5-10	2741	2989.65	3103.25
10-15	618	819.14	651.61
15-20	178	296.94	204.31
20-25	97	122.68	79.99
25-30	42	54.73	36.16
30-35	20	25.67	18.10
35-40	10	12.49	9.78
40-45	5	6.24	5.61
45-50	1	3.18	3.38
>50	4	3.54	6.67

Table 2.2 confirms that even in the upper tail, the lognormal distribution fits the data more accurately than the inverse Gaussian. Although the lognormal underestimates the frequency of observations lying between 20 and 40 ppb, the inverse Gaussian overestimates these frequencies by an even greater margin.

2.4: Conclusion

The work carried out in this chapter has confirmed that the lognormal distribution does not fit the frequency distribution of observed SO_2 concentrations from ITE Bush in 1990 particularly well.

However, the inverse Gaussian distribution has been shown to not be able to improve on this, as the fit has been shown to be even less acceptable, with the values of the lack of fit statistics far more extreme than those for the lognormal. The distribution was fitted by least squares only, as given the very poor fit , the improvement gained by using an alternative such as minimising the Anderson - Darling test statistic will not be sufficient to affect the conclusion that the inverse Gaussian is unsuitable.

Chapter 3: Modelling Instrumental Variation

3.1: Introduction

In many environmental studies today, data are obtained automatically by instruments recording 'continuously', which may mean either point readings at frequent time intervals, or aggregated levels over short time periods. This can lead to an overwhelming volume of data if the observations continue over long periods. Frequently, because of this, the values recorded may be at longer time intervals, with perhaps as many as 100 consecutive observations taken and summarised as an average or summed value and then recorded.

With instruments recording in this manner, the calibration may change over the time periods considered. Ideally, the instrument will have been recalibrated at regular intervals, leaving a gap in the recorded data. The data between successive calibration times may then have been adjusted by a linear transformation to compensate for the instrumental drift observed.

The 'raw data' made available for analysis are the end results of these processes.

Another problem that can occur with data of this type is the presence of outliers caused by a sudden change in the recorded values for only one or two consecutive observations.

If a time series model of the variable is included in the analysis of the data, it is particularly important to understand the structure of this 'raw data'. This chapter is intended to draw attention to the problem, to demonstrate its potential for distorting conclusions, and to indicate one possible method of attack.

3.2: Methods

The data analysed in this chapter are measurements on levels of ground level ozone, measured in parts per billion (ppb), at ITE, Bush Estate, Midlothian, as a small part of a much wider project studying ozone levels (Fowler & Cape (1982) and Smith, Fowler & Cape (1989)). This data set is unusual in that readings are available from two separate instruments, which are operating over different time spans. One (MLO3) is present more or less continuously, and the other (AAO3) is present when not required for fieldwork elsewhere. This provides the opportunity to study the differences between the readings on the two instruments. According to the manufacturer's specifications, these are said to have slightly different characteristics, the AAO3 having better accuracy at low levels of ozone, and a lower detection limit. The MLO3 gives readings every 12 seconds, and the AAO3 every 7 seconds, but for both machines the data provided are averages over 15 minutes of the higher frequency records.

Data were selected from periods when both machines recorded more than 1000 consecutive observations. To investigate the structure of the 'raw data' in this case, the differences between the readings given by the machines at each time point were calculated for the periods when both were operating, and these data sets have been analysed using time series methods. Four such data sets from 1988 have been selected for presentation in this chapter.

3.3: Results

In fig. 3.1 the differences between the 2 readings are plotted as a time series. These plots highlight a problem in this data set with outliers. The spikes that can be seen in the time series plots are caused by the reading from one machine 'jumping' up for one reading and then returning to its previous level. Both analysers exhibit this behaviour with roughly the same frequency. These spikes are thought to be a feature of the data recording device, and not true observations on ozone levels for the following reasons:

1. The 'jump' occurs on only one of the two machines at any one time. If there had been an actual increase in ozone levels over even such a short time span, one would expect both machines to have recorded it.

2. An examination of data for other gases present (nitrogen oxides, sulphur dioxide) did not indicate that an increase in ozone levels had occurred as the level of ozone observed is linked to levels of other gases present in the atmosphere. If the spike were a true observation, one would expect a corresponding spike to occur in the records for the other gases at the same time. For details of the interactions between different pollutants see Fowler & Cape (1982) and Smith, Fowler & Cape (1989).

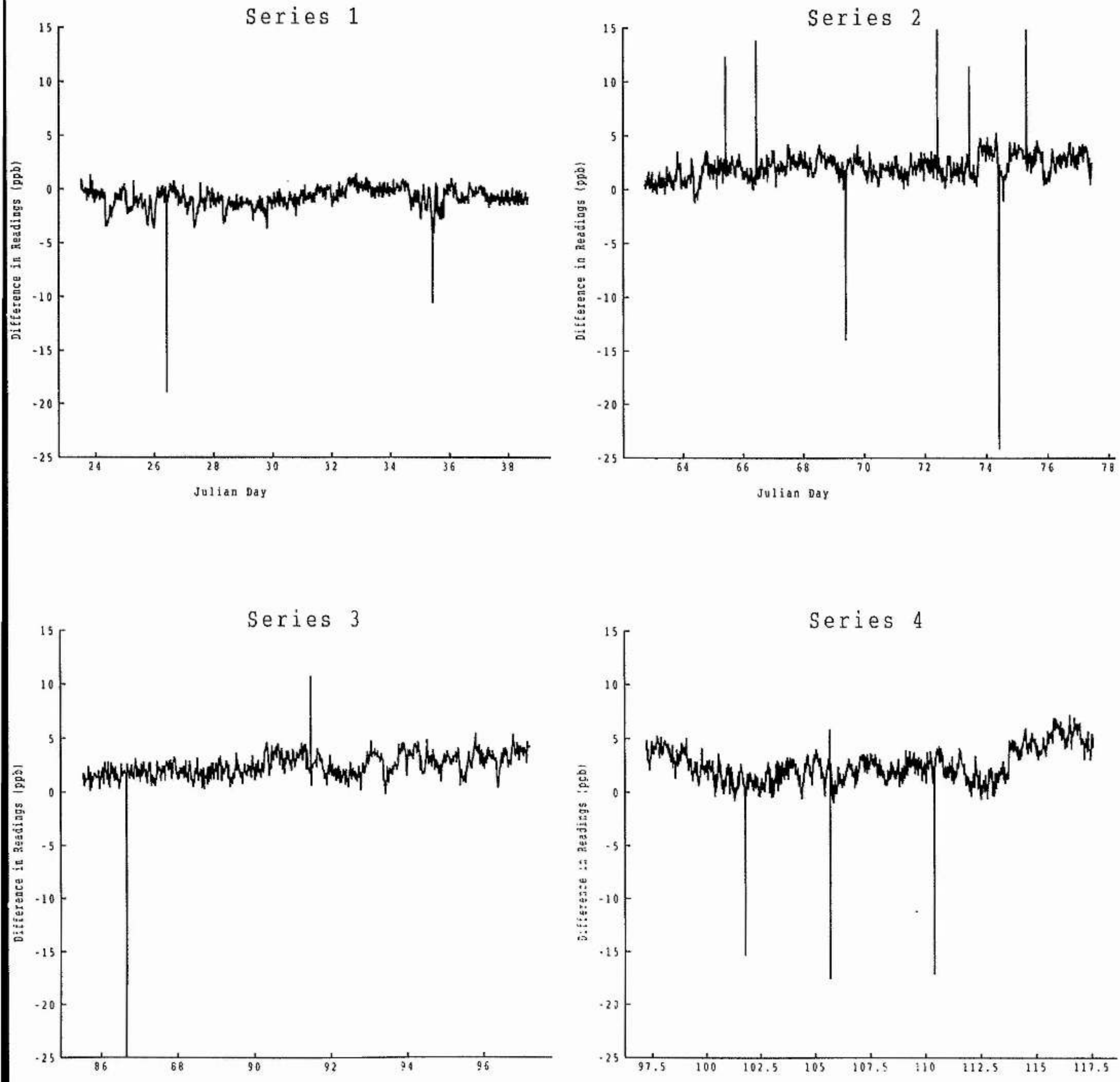
For the remainder of this work, these spikes were treated as missing observations and were replaced using a simple average of the observations immediately before and after the spike.

A spike at time t on machine 1 has been replaced by the value:

$$\frac{(X_{1,t-1} - X_{2,t-1}) + (X_{1,t+1} - X_{2,t+1})}{2} + X_{2,t}$$

Where $X_{i,t}$ represents the reading on machine i at time t. (i=1,2)

Fig 3.1: Differences between readings on ozone analysers



3.4: Time Series Analysis of Differences

Model Selection

The autocorrelation and partial autocorrelation sequences for the data were calculated (with the spikes removed), these are shown in figs 3.2-3.5. The PACF decays to near zero after lag 1 or 2 while the ACF shows a more gradual decay. After differencing the series once, both the ACF and PACF exhibit a rapid decay after lag 1 or 2. Plots of the data after differencing show no obvious trends in level or variability, implying stationarity. There are also no obvious short or long term seasonal effects. These facts point to the use of one of three different Box-Jenkins type models (Wei 1990), which are described below:

1. IMA(1,1):

$$Z_t = Z_{t-1} + a_t - MA(1)\, a_{t-1} \qquad \text{-(1)}$$

(where Z_t is the difference at time t, and a_t is a random shock at time t)

2. IMA(1,2):

$$Z_t = Z_{t-1} + a_t - MA(1)\, a_{t-1} - MA(2)\, a_{t-2} \qquad \text{-(2)}$$

3. ARIMA(1,1,1):

$$W_t = AR(1) W_{t-1} + a_t - MA(1)\, a_{t-1} \qquad \text{-(3.1)}$$

where

$$W_t = Z_t - Z_{t-1} \qquad \text{-(3.2)}$$

Fig 3.2: ACF and PACF for series 1, before and after differencing

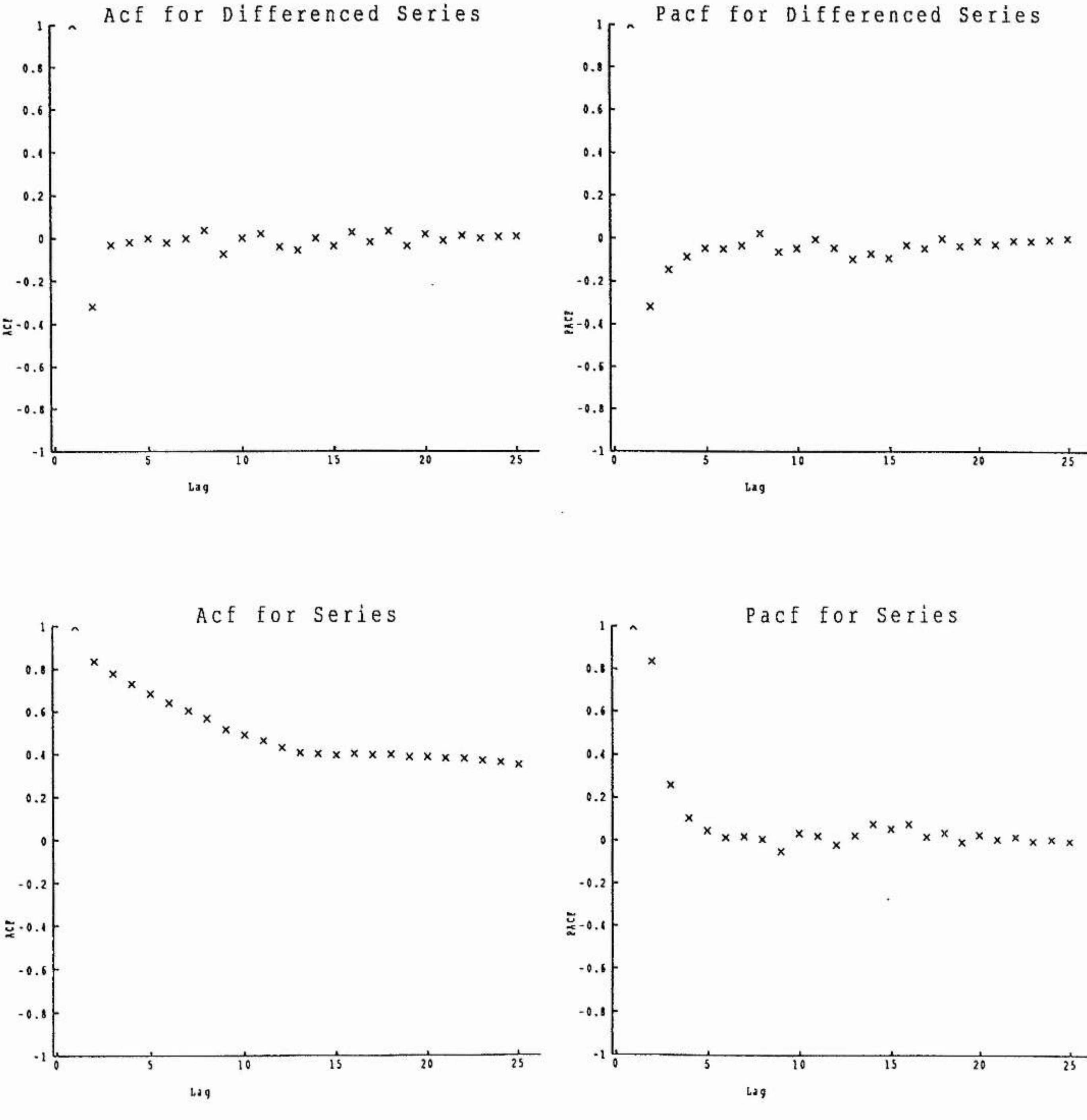


Fig 3.3: ACF and PACF for series 2, before and after differencing

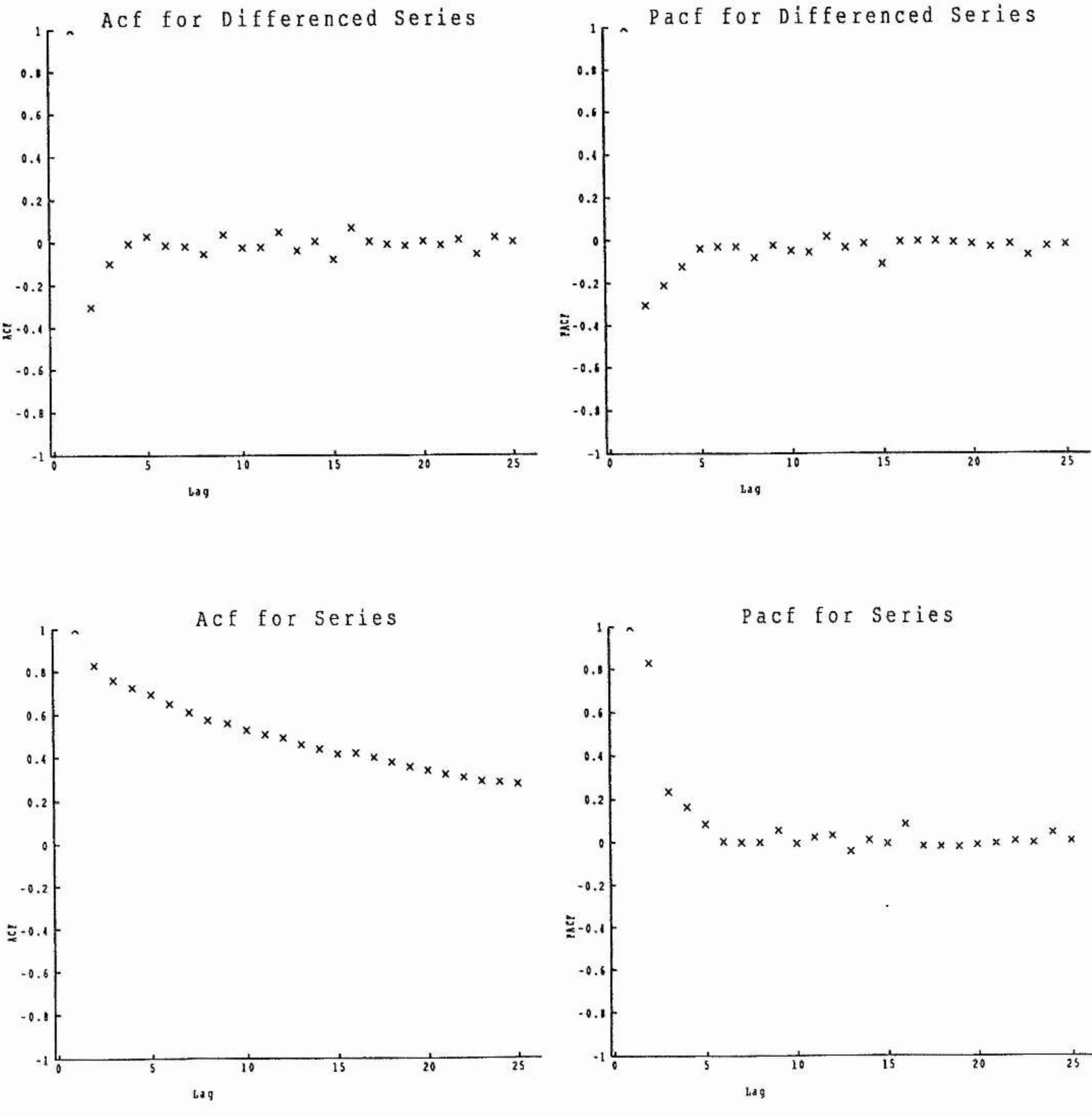


Fig 3.4: ACF and PACF for series 3, before and after differencing

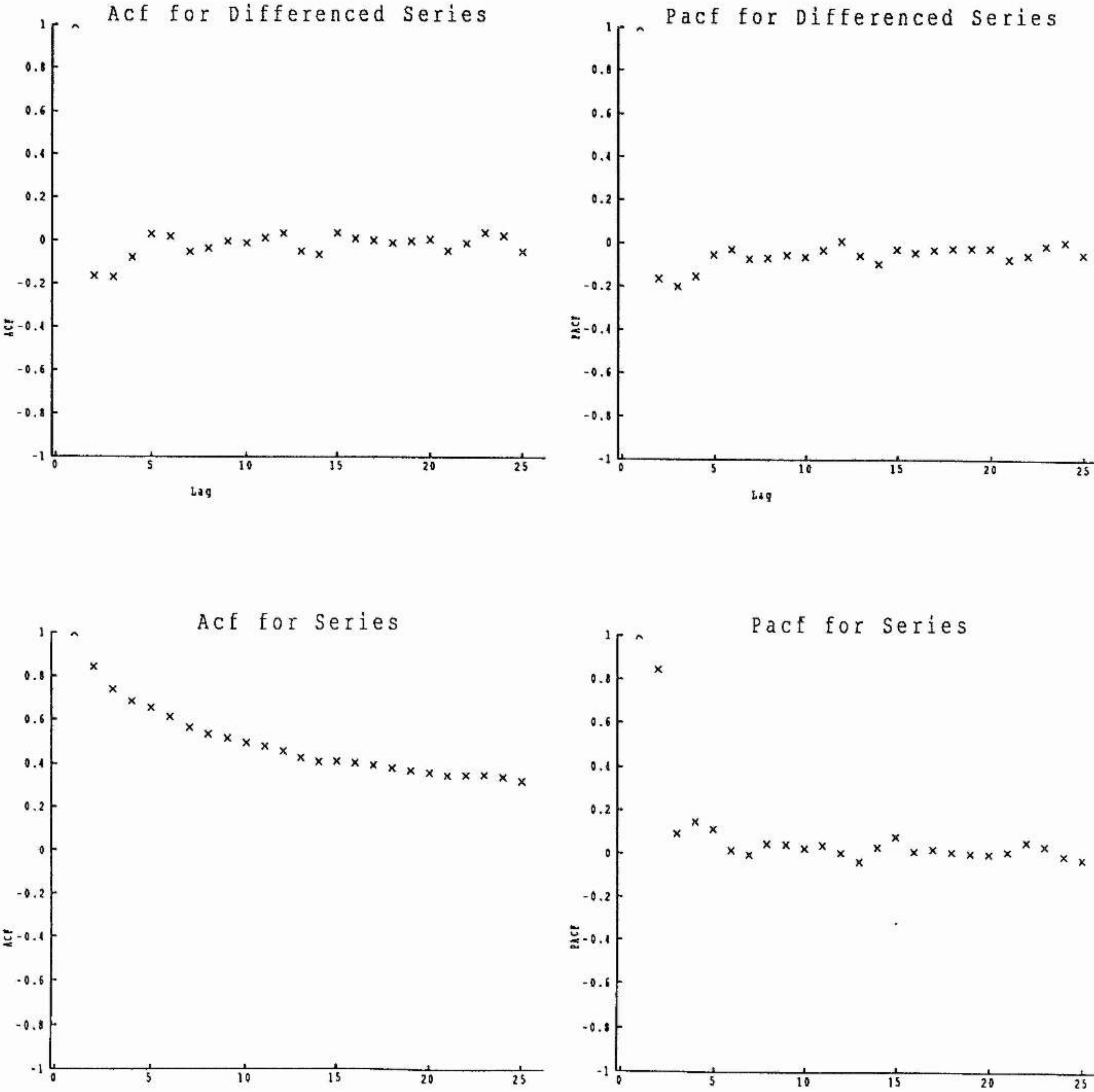
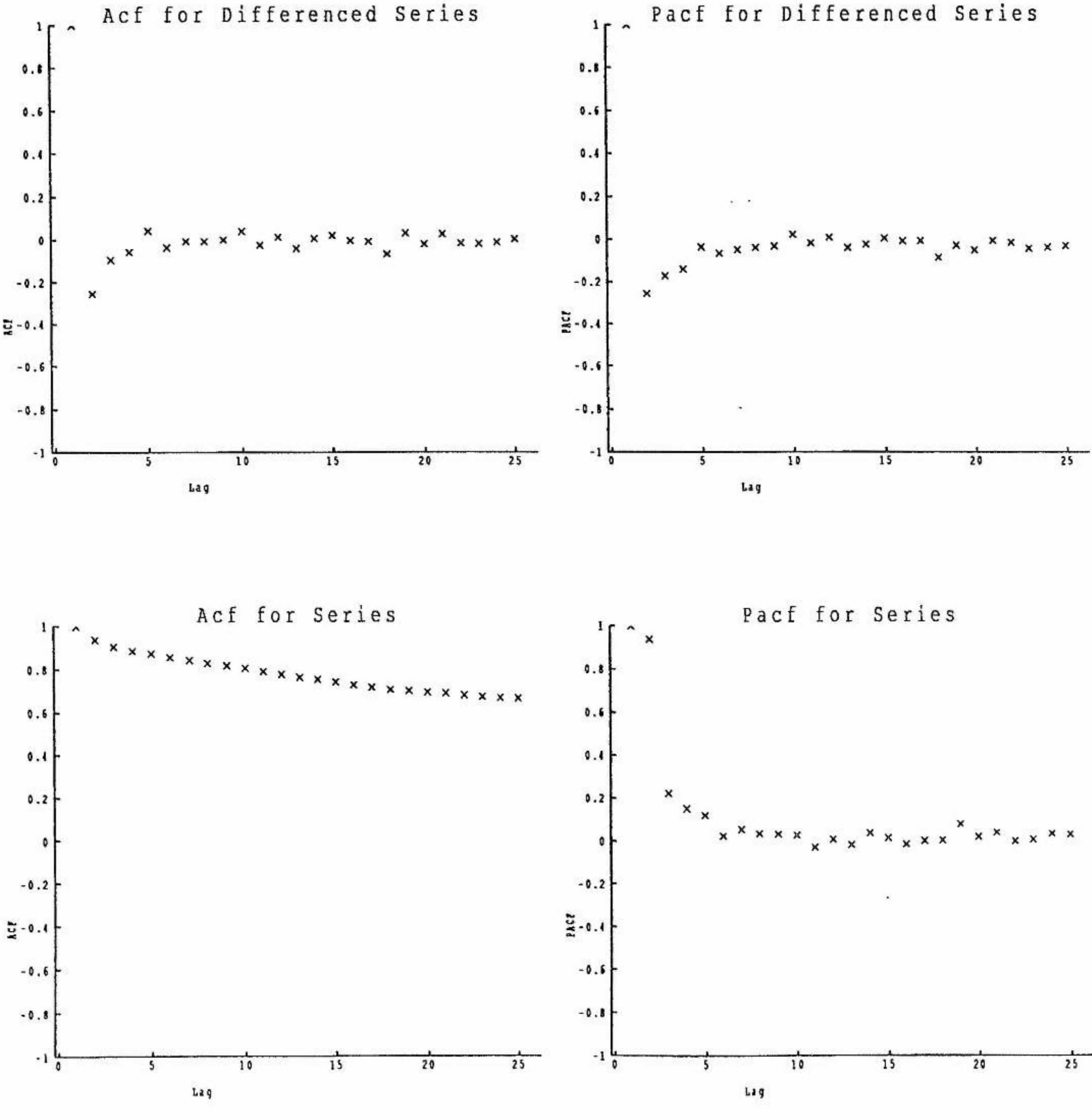


Fig 3.5: ACF and PACF for series 4, before and after differencing



Acf for Differenced Series

Pacf for Differenced Series

Acf for Series

Pacf for Series

Model Fitting

Table 3.1 contains the estimates of the parameters for each of the three models indicated as possibilities above for the four data sets considered here, both with the spikes present and removed. The t-ratio also quoted is an indication of whether or not the parameter is required in the model, with a low t-ratio leading to the conclusion that the term with that parameter should not be included. For details see Wei (1990).

TABLE 3.1 - Estimates of parameters and t-ratios

	IMA(1,1)		IMA(1,2)				ARIMA(1,1,1)			
Series	MA1	t-ratio	MA1	t-ratio	MA2	t-ratio	AR1	t-ratio	MA1	t-ratio
1 no spikes	0.40	16.83	0.39	14.71	0.06	2.26	0.18	2.97	0.56	10.96
1 spikes	0.63	31.24	0.60	22.89	0.07	2.54	0.12	2.89	0.71	25.22
2 no spikes	0.45	19.06	0.40	15.28	0.12	4.52	0.27	5.15	0.67	16.85
2 spikes	0.80	51.31	0.77	28.91	0.06	2.07	0.08	2.42	0.84	48.91
3 no spikes	0.30	10.12	0.27	9.17	0.23	7.80	0.69	26.65	0.94	87.50
3 spikes	0.89	67.54	0.84	28.54	0.07	2.43	0.09	2.91	0.92	76.41
4 no spikes	0.38	18.17	0.34	15.26	0.14	6.18	0.36	7.76	0.04	19.47
4 spikes	0.73	47.79	0.73	32.19	0.01	0.27	0.01	0.30	0.74	35.66

For all of the data sets used, both with and without spikes, the first MA parameter is seen to be highly significant, whilst the second MA parameter is just significant with the spikes removed, and less significant in all cases with the spikes present. The AR parameter is also just significant when applied, except for series 3 with the spikes removed, when it is highly significant. The spikes can be seen here to cause a significant change in the estimates of each of the parameters fitted.

Table 3.2 contains the values of a test of normality applied to the residuals from each of the models essentially equivalent to the Shapiro Wilk test, based on the correlation of the data and their 'normal scores' - for details see Filliben (1975). The critical value for this test for n=100 is 0.994. The value of this test statistic does not change significantly across the three models considered, but does across different data sets. The residuals from the models fitted to the data sets with the spikes not removed can definitely not be regarded as normally distributed, whilst with the spikes removed only the residuals from data sets 1 and 4 are questionably normal. Normal probability plots and histograms of these residuals suggest that they come from a symmetric distribution, with slightly heavier tails than the normal.

TABLE 3.2 - Test of normality on residuals

Series	IMA(1,1)	IMA(1,2)	ARIMA (1,1,1)	Number of observations
1 no spikes	0.992	0.991	0.991	1456
1 spikes	0.750	0.748	0.74	1456
2 no spikes	0.998	0.998	0.998	1414
2 spikes	0.690	0.686	0.685	1414
3 no spikes	0.998	0.997	0.998	1123
3 spikes	0.571	0.564	0.562	1123
4 no spikes	0.991	0.992	0.992	1954
4 spikes	0.711	0.710	0.710	1954

Table 3.3 contains the values of the sum of squared residuals and the mean square of the residuals after fitting each of the three models. The overall lack of fit does not change significantly across the three types of model considered, but the lack of fit is much greater with the spikes present.

TABLE 3.3 - SS and MS residual after model fitting

Series	IMA(1,1)		IMA(1,2)		ARIMA (1,1,1)	
	SS	MS	SS	MS	SS	MS
1 no spikes	315.83	0.22	314.79	0.22	314.55	0.22
1 spikes	811.39	0.56	808.28	0.56	808.01	0.56
2 no spikes	455.23	0.32	449.28	0.32	449.16	0.32
2 spikes	2385.03	1.69	2378.39	1.69	2377.65	1.69
3 no spikes	339.43	0.30	323.29	0.29	317.35	0.28
3 spikes	2680.61	2.39	2666.0	2.38	2664.41	2.38
4 no spikes	526.38	0.27	516.13	0.27	514.69	0.26
4 spikes	1786.04	0.91	1785.97	0.92	1785.97	0.92

The ACF and PACF sequences for the residuals from each of the models fitted were also examined, and no significant autocorrelation structure was observed.

From the information given, it can be seen that the spikes have a large effect on model fitting and selection, and that their removal is necessary. The best choice of model can also be seen to be either the IMA(1,1) or IMA(1,2) model, but the overall lack of fit is not badly affected by the choice of model.

3.5: Discussion

The IMA(1,1) model:

$$Z_t = Z_{t-1} + a_t - \mathrm{MA}(1)\, a_{t-1}$$

(where Z_t is the difference at time t, and a_t is a random shock at time t)

would be an intuitively reasonable model for the series. To see this, write the model for the Z's explicitly as a function of past and present a-values:

$$Z_t = a_t + (1-\theta)a_{t-1} + (1-\theta)a_{t-2} + \ldots + (1-\theta)a_{-m} - \theta a_{-m-1}$$

where -m is earlier than time t=1, at which point we first observed the time series, and θ is the MA(1) parameter in the model. Since we are assuming that $-m<1$ and $0<t$, we may usefully think of Z_t as being an equally weighted accumulation of a large number of white noise values.

In terms of the ozone analysers, if we regard each white noise value as being the random drift in calibration between the two machines during the fifteen minute period, the value of Z_t being an accumulation of these would seem to be a reasonable model for the data (i.e. a sum of all these random drifts since the analysers were turned on together).

The observations recorded are not, in fact, the data produced by the analysers directly. The fifteen minute readings given are averages of readings given at intervals of seven second readings on one machine and twelve second readings on the other. This fact may also affect the choice of model for the data given.

The theory of temporal aggregation of the ARIMA process (Wei 1990) states that if a disaggregated series (denoted by z_t) is aggregated to form the series Z_t, the m-period nonoverlapping aggregates of z_t, by summing every m successive z_t's, then the following results hold:

1. If z_t follows an IMA(d,q) model, then Z_t follows an IMA(d,N_0) model, where N_0 is defined by:

$$N_0 \leq q^* = [d+1+(q-d-1)/m],$$

and [x] is used to denote the integer part of x.

2. If z_t follows an ARIMA(p,d,q) model, then as m -> ∞ , the limiting model for the aggregates Z_t exists and is the IMA(d,d) process, independent of p and q. A similar result holds if the disaggregate series follows a seasonal ARIMA(p,d,q)x$(P,D,Q)_s$ model, leading to an IMA(D+d,D+d) process.

These results give a possible justification for selecting the IMA(1,1) model for the data as follows:

The data given are aggregates of a large number of observations, with m being equal to 75 for the MLO3 data and 129 for the AAO3 data.

If we assume that the disaggregated series for the data considered follows an IMA(1,q) model, then result 1 states that the aggregated series would then follow an IMA(1,1) model if q=1; or either an IMA(1,1) or IMA(1,2) model if $2 \leq q < m+2$. If we assume that the disaggregated series follows any ARIMA(p,1,q) model, or any ARIMA(p,d,q)x$(P,D,Q)_s$ model with p+D=1, and that m is large enough for result 2 to hold, then the aggregated series would follow an IMA(1,1) model. Thus it can be seen that a large number of possible models for the disaggregated series lead to an IMA(1,1) model as a model for the aggregates.

Unfortunately, only a very small amount of data from the disaggregated series is available, and if some can be obtained some useful work in this direction could be done.

It is also possible that the data could be transformed in some way to improve the fit to the model, but some transformations have been attempted without much success. The log transformation cannot be applied to the differences directly as the data are equally positive and negative, and some simple power transformations have been attempted. It is thought that it may prove useful to transform the readings from the analysers, and then to work on the differences between the transformed data.

3.6: Conclusion

The work outlined in this chapter indicates that one of the models considered may prove useful when analysing data of this type. Although the lack of fit of the model could be improved by either transforming the data in some way to induce normality, as the correlation structure appears to be well modelled, or using an IMA type model with a non-gaussian random shock distribution. Also, a study of the disaggregated series could prove highly useful in finding the type of model required.

The spikes observed in the data are also thought to be highly important, for two reasons:

First, it is seen that the analysis described in this chapter is affected if the spikes are ignored, the choice of model for a particular data set may change, and the fit of the model is also very much affected.

Second, much analysis is carried out on extreme value statistics in this field, as the maximum values of pollutant data are those which are often the most important to the analyst. The spikes observed in the data sets analysed here would often be recorded as the daily maxima, thus affecting an extreme value analysis greatly. If other data sets exhibit these, and they are not detected and removed if they are not true observations, the effect on modelling could prove disastrous.

Chapter 4: Modelling Monthly Average Ozone Curves

4.1: Introduction

A common feature of analyses of ozone monitoring data is the average diurnal cycle. This average cycle is calculated either as the average over a year or, in some cases, as the average over a few months. For example, in the PORG interim report (1987), a variety of diurnal cycles are displayed. Some are the averages over April to September, some are averages over a whole year, and one figure displays the six cycles for the months July to December separately. However, the only discussion of how the cycle varies over a year is a statement that there is generally a more pronounced diurnal cycle observed in the summer.

It is known that observed ozone levels depend largely on the weather (PORG 1987 & 1993) and that the distribution of weather types varies over a year. These facts should lead to changes in the diurnal behaviour of ozone at different times of the year. This chapter will investigate the variation in the average diurnal cycle over a year.

To carry out this investigation, it is proposed to calculate the average diurnal cycle for each month of a year and model these monthly averages. A month has been chosen for several reasons: it will contain a variety of weather types as in the UK weather patterns persist for 4-5 days on average; it will be a long enough period to not be badly affected by any short term freak weather patterns; a month contains sufficiently many individual days to meaningfully calculate an average cycle, and will divide the year into sufficiently many sections to be able to evaluate any changes in the diurnal behaviour of ozone over a year.

4.2: Methods

The data analysed in this chapter are measurements on levels of ground level ozone, measured in parts per billion (ppbv), at ITE, Bush Estate, Midlothian during 1988. The 'raw data' are the 15 minute averages from the MLO3 machine considered in the previous chapter, with all the known 'spikes' removed.

To calculate the average diurnal cycle over a specific period, all the observations available at each of the 96 15 minute time points in a day have been averaged separately:
If O_{ij} represents the ozone level at time j in day i, (i=1...N , j=1...96), the level of the average curve for these N days at time j, A_j is calculated as:

$$A_j = \frac{1}{N}\sum_{i=1}^{N} O_{ij}$$

The average diurnal cycle for the whole of 1988 is shown in fig 4.1. The average diurnal cycles for each of the twelve months of 1988 are shown in fig 4.2.

These average curves differ in overall level (mean), spread (variance) and shape. The question of shape is dealt with below. It would be possible to standardise the data by mean and standard deviation to consider only variation in shape, but for modelling purposes these would have to be recorded and used as parameters. Indiscriminate standardisation can also adversely affect the results of an analysis. For example Matz(1992) carried out a short investigation of the variation between monthly averaged curves from Mexico city in 1990. As the first stage of his analysis he standardised each day by dividing all its values by the daily maxima. As the maxima invariably occur during the daytime, two nights with identical behaviour will be very different if the previous daily maxima were very different. His conclusion that the main source of variation is the changes in behaviour at night is thus suspect, as this may simply be a product of his standardisation method.

Fig 4.1: Bush 1988 Average Ozone diurnal curve

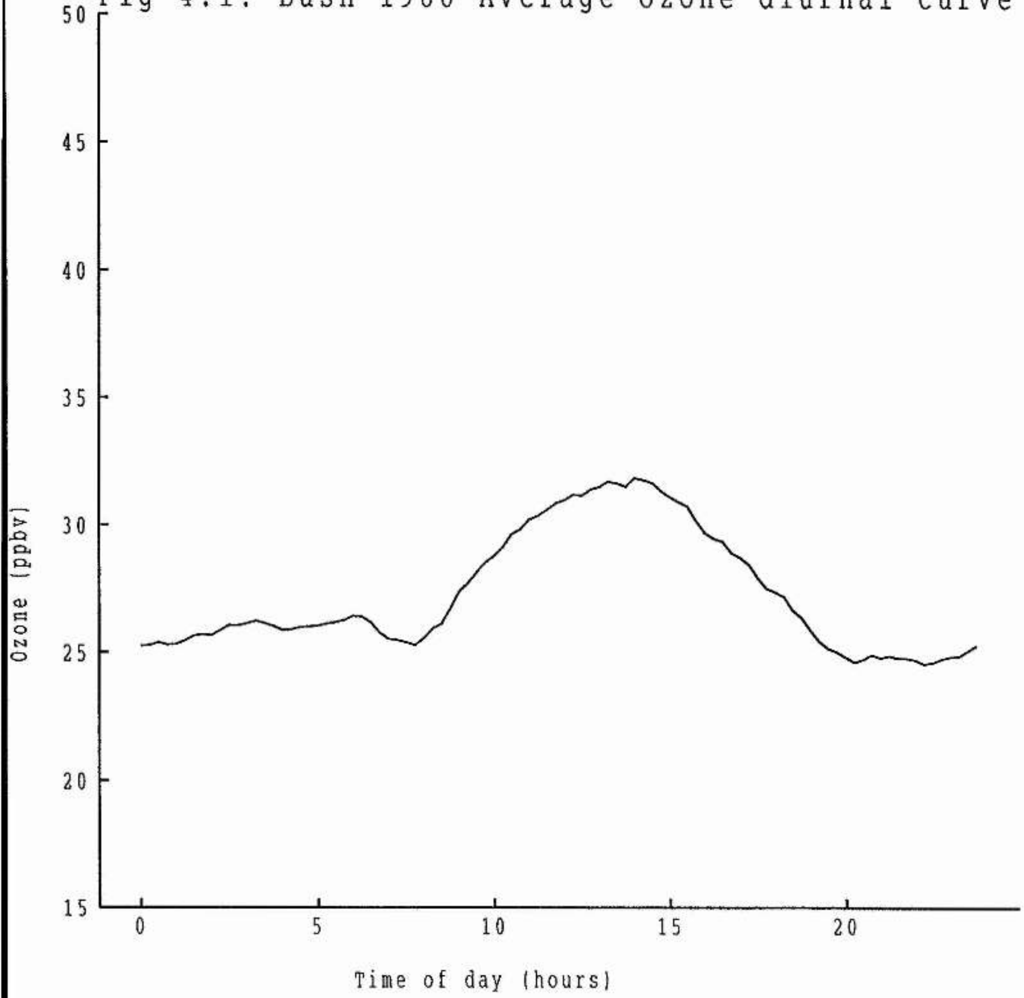


Fig 4.2 (part 1): Monthly average ozone diurnal cycles for 1988

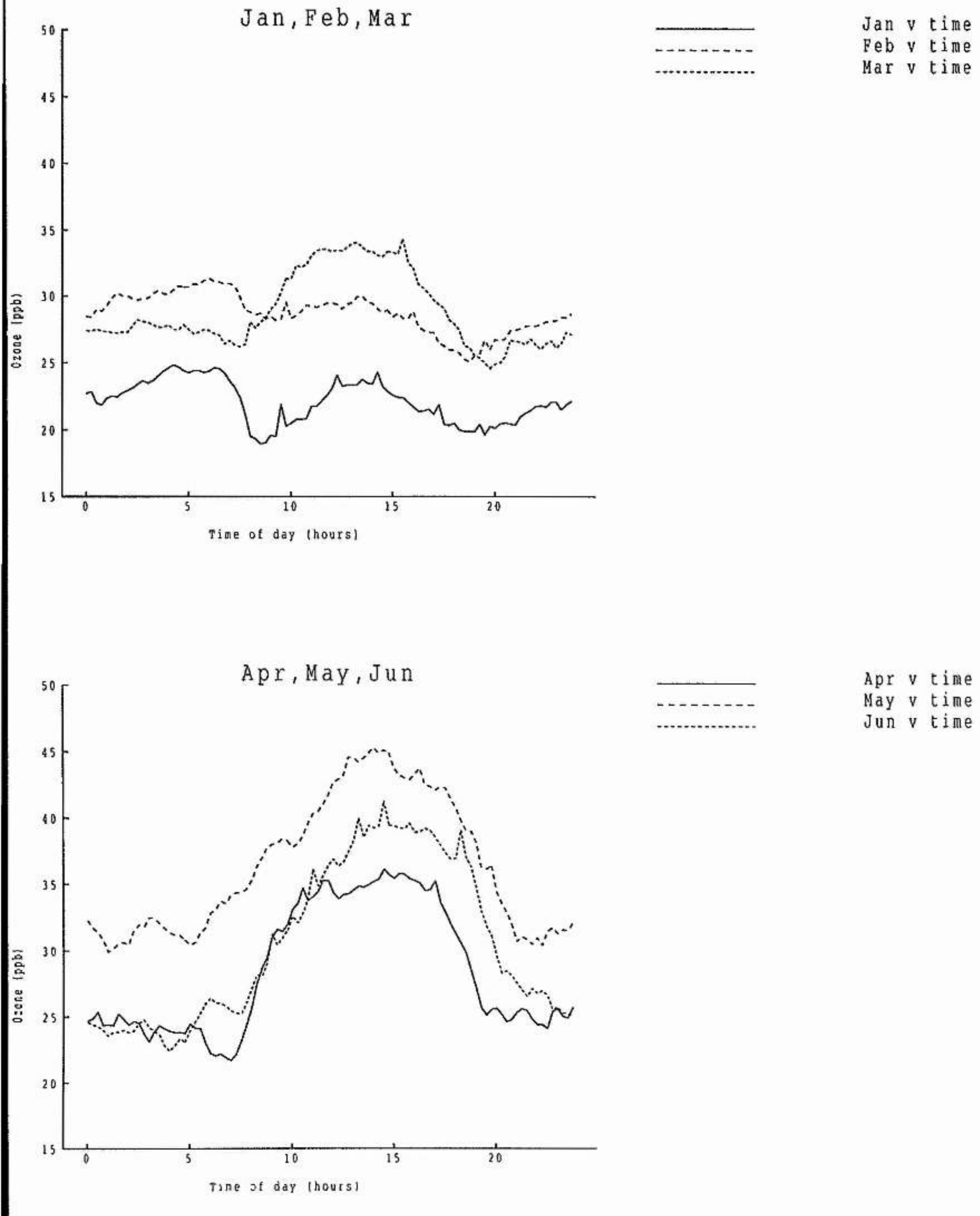
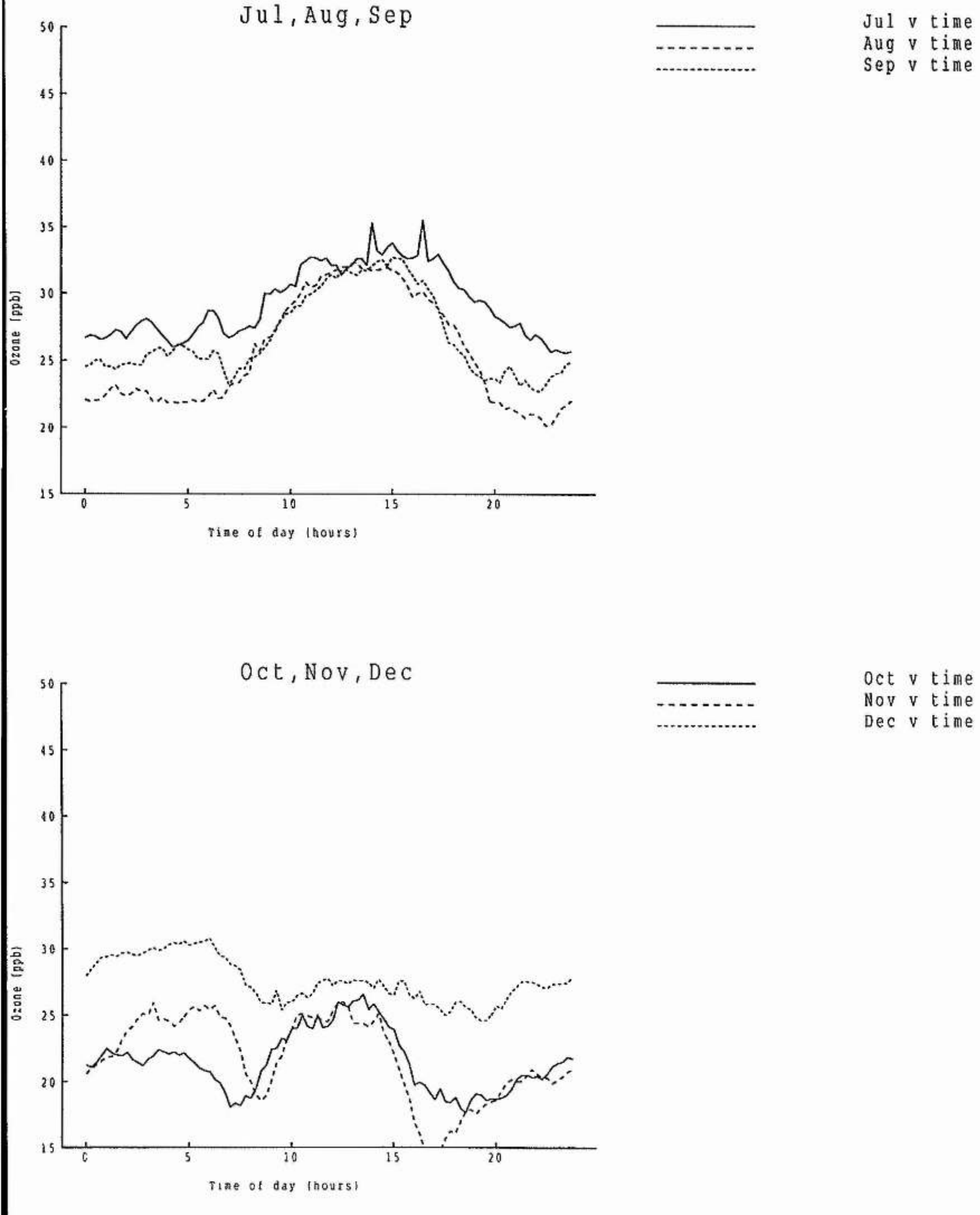


Fig 4.2 (part 2): Monthly average ozone diurnal cycles for 1988



Time Distortion

Fig 4.2 demonstrates that the diurnal cycles have differing shapes for each of the twelve months. They all have a 'hump' starting at between 5 and 8 am and finishing between 4 and 8 pm. This section of the curve represents the daytime variation in ozone levels and is driven by a combination of factors including sunlight, temperature and wind speed. These factors depend more on the position of the sun in the sky than clock time. This suggests using the time of day relative to the sun and not clock time as the time used on the x-axis.

To do this a transformation of the time axis is required for each day of the year (thanks to C.D.Steele for help with this transformation) which is carried out as follows:

1. For each day of the year where observations are available calculate declination of the sun, δ, using the following equation:

$$\sin \delta = \sin \varepsilon \sin \frac{360\ d}{365.25}$$

$\varepsilon=23.45^{o}$

d=days since March 21st

(d negative 1st Jan - 20th March)

2. Once this has been carried out, use the value of δ to calculate transformed times t' for each time point t during that day where an observation is given. For each time point t calculate transformed times t' as follows:

From the time t, in hours, calculate angular time A (degrees):

$A = 15*t$ t=time in hours

Then calculate the transformed angular time A' using the following equation:

$$\tan A' = \frac{-\cos\phi \sin A}{\sin\phi \tan\delta - \cos\phi \cos A}$$ Where ϕ = latitude.

The transformed time in hours t' is then given by:

$t' = A'/15$

The transformed times t' have been calculated for every time point t at which an ozone reading is available. This results in a set of daily ozone curves plotted on a time axis where sunrise, sunset, midday and midnight occur at the same transformed time for every day of the year.

However calculating the average curves for each month is now complicated by the fact that the points on the curves are not given at the same time points on the transformed time axis for each day. To overcome this problem, values have been interpolated from each of the daily curves on the transformed time axis so that values are available at the same, equally spaced, transformed times for each day. Simple linear interpolation was used between two nearest points. The errors introduced into the data by this procedure will be small, as the maximum change in ozone levels between two successive 15 minute time points is small (97% of such changes are less than 4ppb in the data set used here).

The new monthly averages have then been calculated from these transformed daily curves. Fig 4.3 shows these new average curves. The daytime 'hump' now starts and finishes at very much the same time for each of the twelve curves.

Fig 4.3 (part 1): Monthly average diurnal cycles (Transformed time)

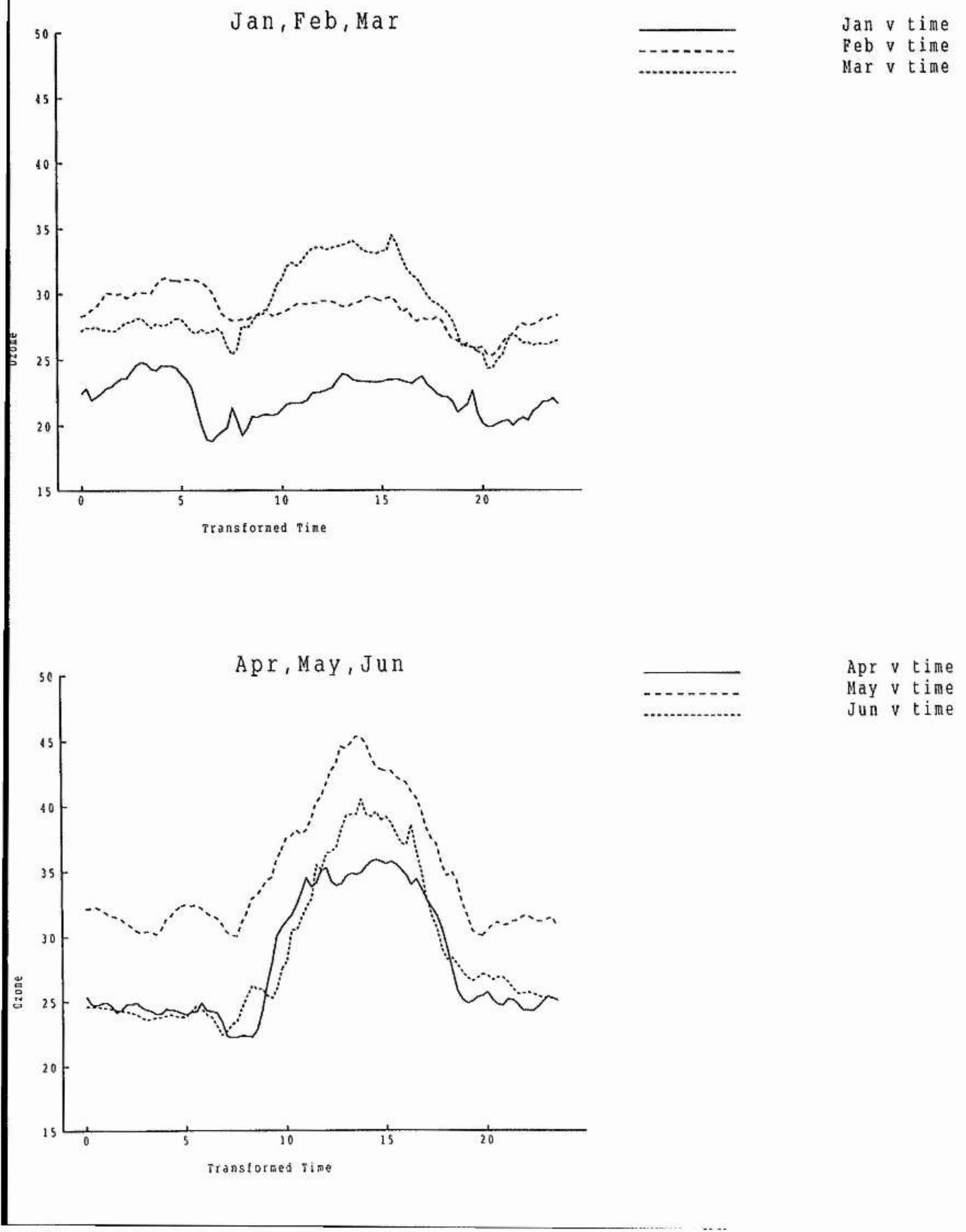
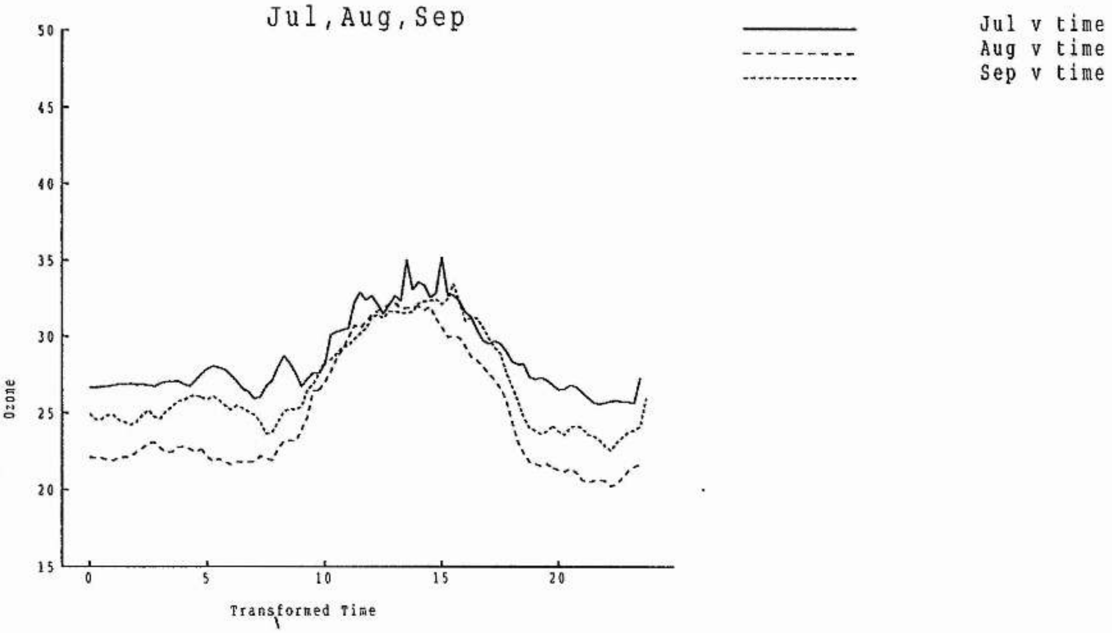
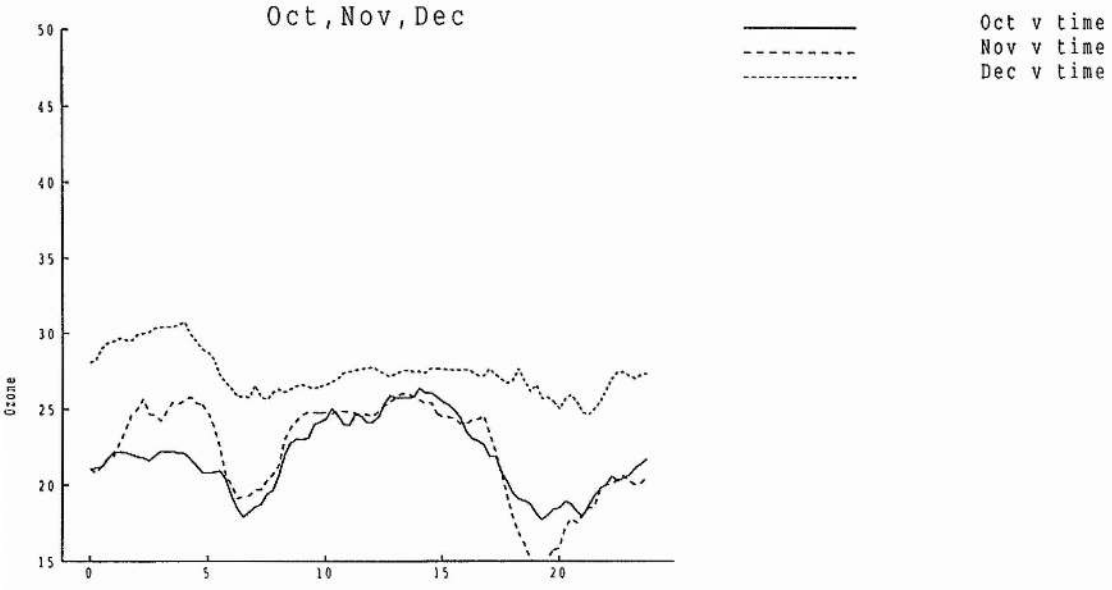


Fig 4.3 (part 2): Monthly average diurnal cycles (Transformed time)





4.3: Model For Ozone Curves

Fig 4.4 shows the same curves as fig 4.3, calculated with the start of each day at 7am on the transformed time axis. 7am transformed time corresponds to 1/24th of the daytime after sunrise. This is when one expects the daytime processes to begin affecting the observed ozone concentration as it will on average take approximately an hour for the night time inversion layer to break up.

The curves plotted this way naturally divide into two sections, representing the daytime and night time variation of ozone separately. Both the daytime and night time sections exhibit variation in spread and level, but the shapes remain very similar from month to month. A model for these data will have to be capable of reproducing this variation in spread and level separately for both sections of the day.

Day and night sections have been modelled separately as polynomials in time of the fifth degree. These have been constrained to be continuous in value and first derivative at the join between day and night (at the centre of the time axis) and at the beginning and end of the day (at the ends of the time axis). Quintics have been chosen as the lowest degree polynomial that will have sufficient flexibility after the constraints are implemented to model the observed variation in the average curves. Each quintic has six parameters (a total of 12 for each curve), which will reduce to four (a total of 8) once the constraints have been implemented. This will provide enough flexibility to model the observed variations in shape, level and variation observed for both the day and night time sections of the curves.

Fig 4.4 (part 1): Monthly average diurnal cycles (Shifted transformed time)

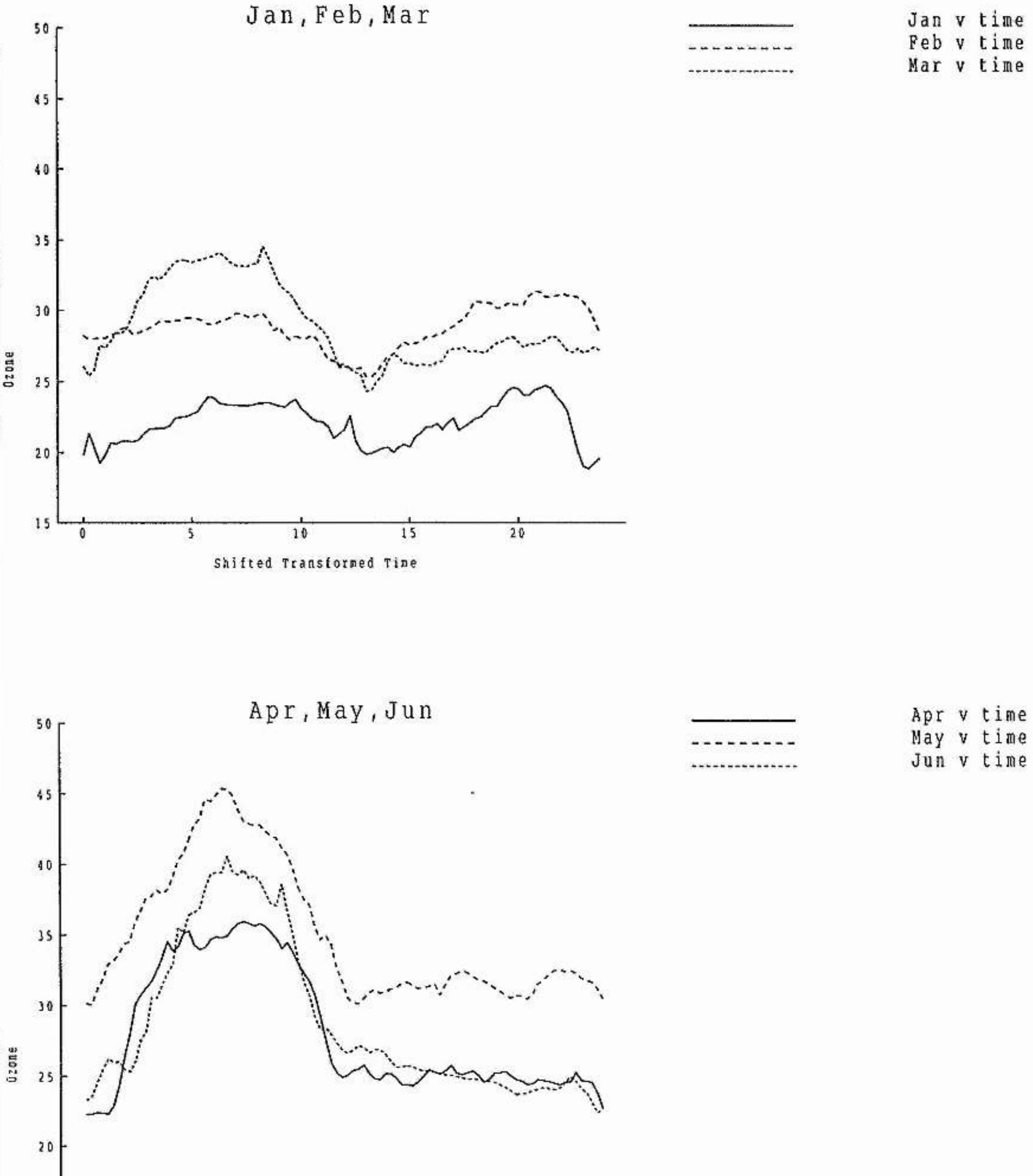
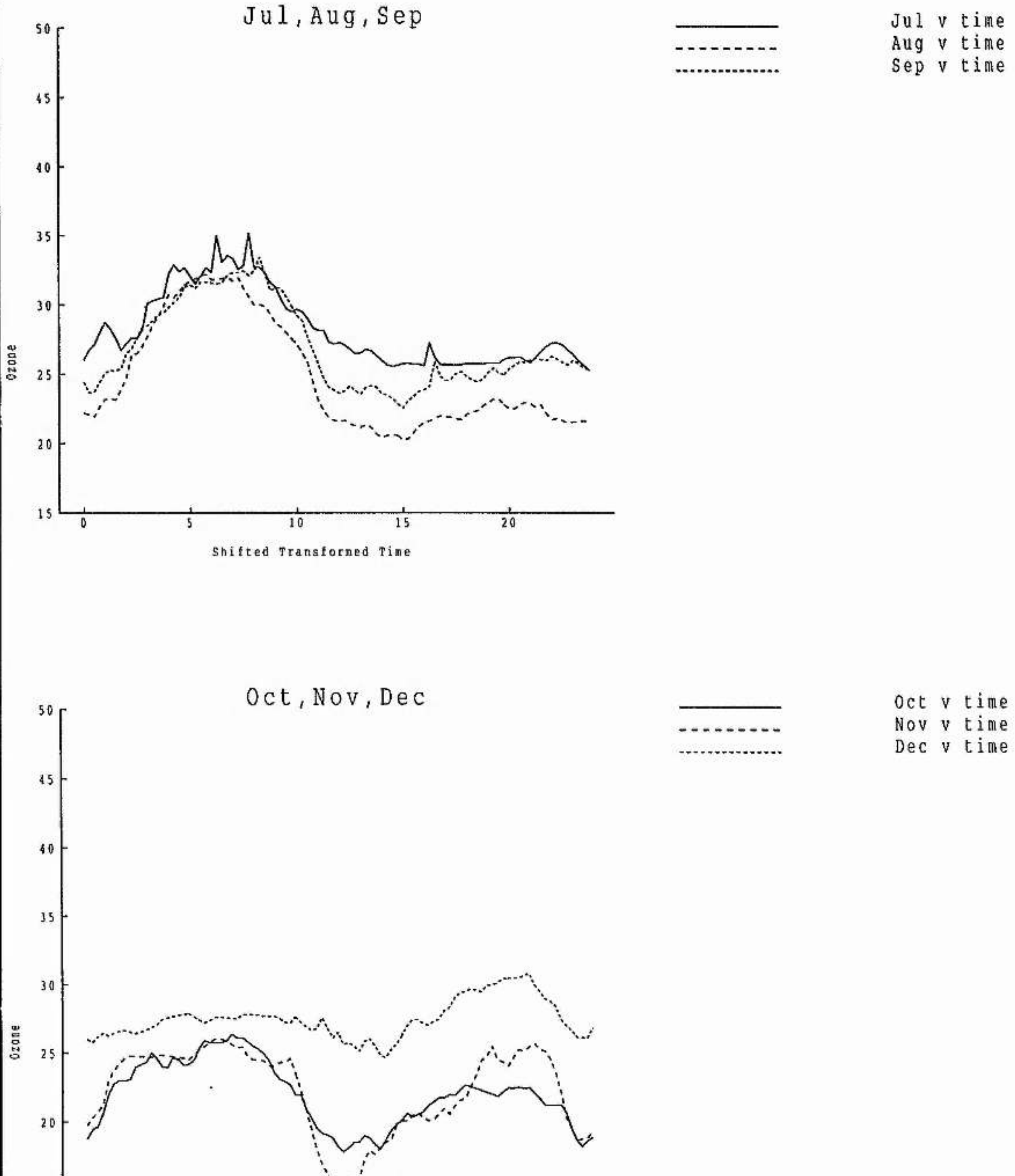


Fig 4.4 (part 2): Monthly average diurnal cycles
(Shifted transformed time)



For algebraic simplicity use $-1 \leq x < +1$ as x variable, -1 is 7am transformed time, and +1 is 7am transformed time the next day. The model can be written as:

$$E[y_{ij}] = \begin{cases} Q_1(x) & -1 \leq x < 0 \\ Q_2(x) & 0 \leq x < 1 \end{cases}$$

Where $Q_1(x)$, $Q_2(x)$ are polynomials of the fifth degree in x with constraints:

$$Q_1(-1) = Q_2(+1) \; ; \; Q_1(0) = Q_2(0)$$

$$Q_1'(-1) = Q_2'(+1) \; ; \; Q_1'(0) = Q_2'(0)$$

Each quintic has six coefficients, twelve in all for each month. The continuity conditions impose four constraints on these, leaving a total of eight independent parameters for each of the twelve monthly curves. The model will thus reduce to a series of linear combinations of eight polynomial functions of x. When the constraints are implemented the model reduces to:

$$E[y_{ij}] = \sum_{k=1}^{8} \theta_{ki} \, f_{kj} \qquad i=1\ldots12, \; j=1\ldots96$$

Where:

y_{ij}: obs in month i, time point j.

Time $x = -1+2/96 * (j-1)$

f_{kj}: polynomials f_k in x at time point j.

θ_{ki}: parameters to be estimated.

(coeffs for polynomial k in month i)

The polynomials f_k are determined by the constraints imposed and are given by:

$$f_1 = 1$$

$$f_2 = (x^3 - x)$$

$$f_3 = (x^4 - 2x^2)$$

$$f_4 = (x^5 - x)$$

$$f_5 = ((x_+)^2 - 0.5x - 0.5x^2)$$

$$f_6 = ((x_+)^3 - 0.5x - 0.75x^2)$$

$$f_7 = ((x_+)^4 - 0.5x - x^2)$$

$$f_8 = ((x_+)^5 - 0.5x - 1.25x^2)$$

$$\text{Where } (x_+)^n = \begin{cases} 0 & x<0 \\ x^n & x\geq 0 \end{cases}$$

The term f_{kj} in the model is f_k evaluated at $x = -1+2/96 * (j-1)$.

This model can be fitted directly to the data, using a least squares algorithm. The parameters are the θ_{ki}, eight for each of the twelve monthly curves. Thus this model requires a total of 96 parameters to model the whole year's monthly variation.

Fig 4.5 shows the twelve curves generated by this model, with the actual data points plotted for comparison. This model fits the monthly averages well. To allow comparison with other modelling approaches that will be considered later, the lack of fit of the model has been assessed by calculating the area between the curve generated by the model for each month and the data curve. Using area between two curves as a measure of the distance between them has various advantages over most standard measures as is discussed in chapter 6 of this thesis. The lack of fit results for this model are displayed in the first row of Table 4.1. The only notable discrepancy is November, but as the curve for November has a more erratic shape than all the other months it is unlikely that any simple curve would be able to improve on the fit.

Fig 4.5 (part 1): Fit by 8 parameter 2-Quintic model

—— raw data
xxxx model fit

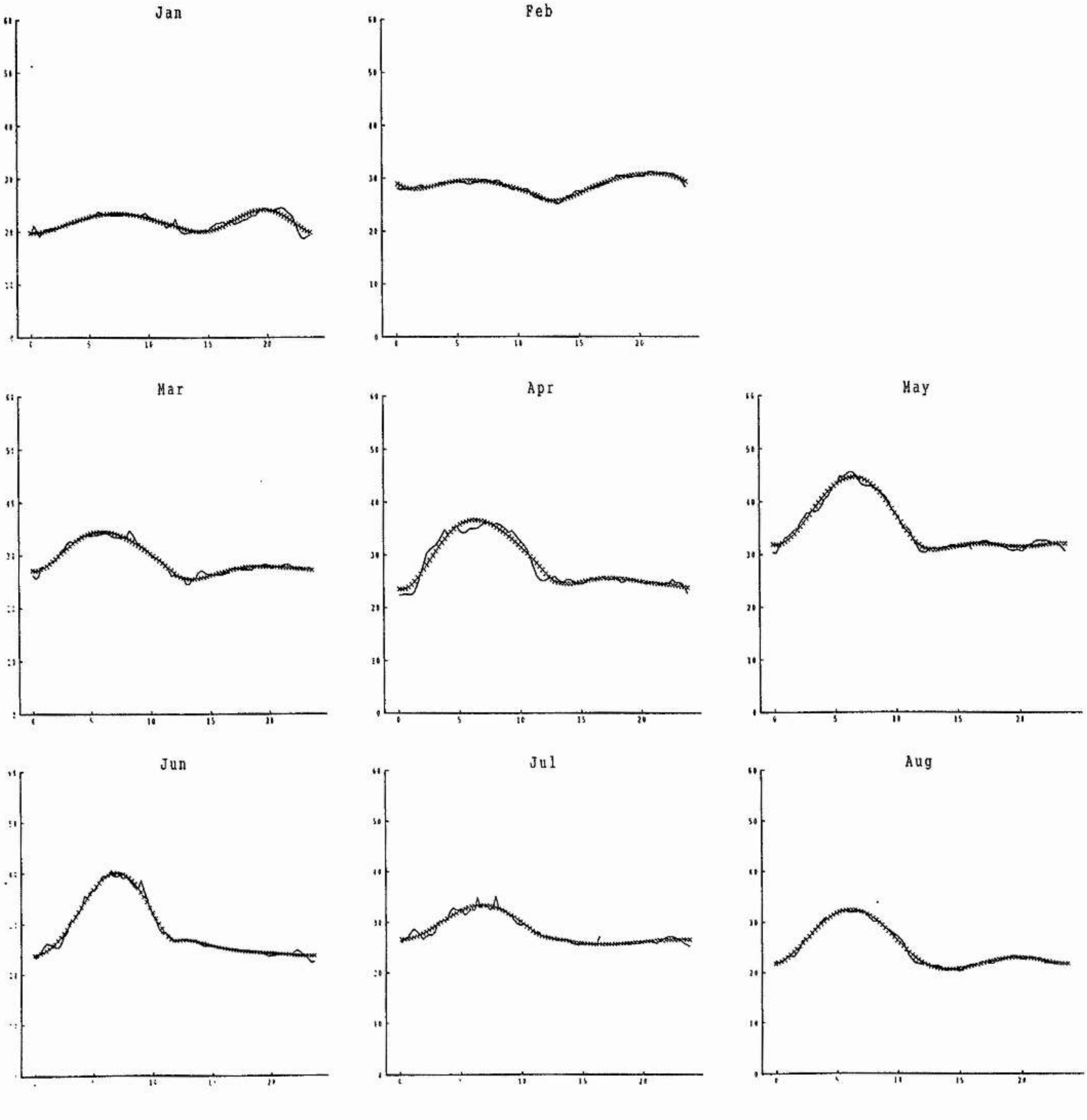


Fig 4.5 (part 2): Fit by 8 parameter 2-Quintic model

—— raw data
xxxx model fit

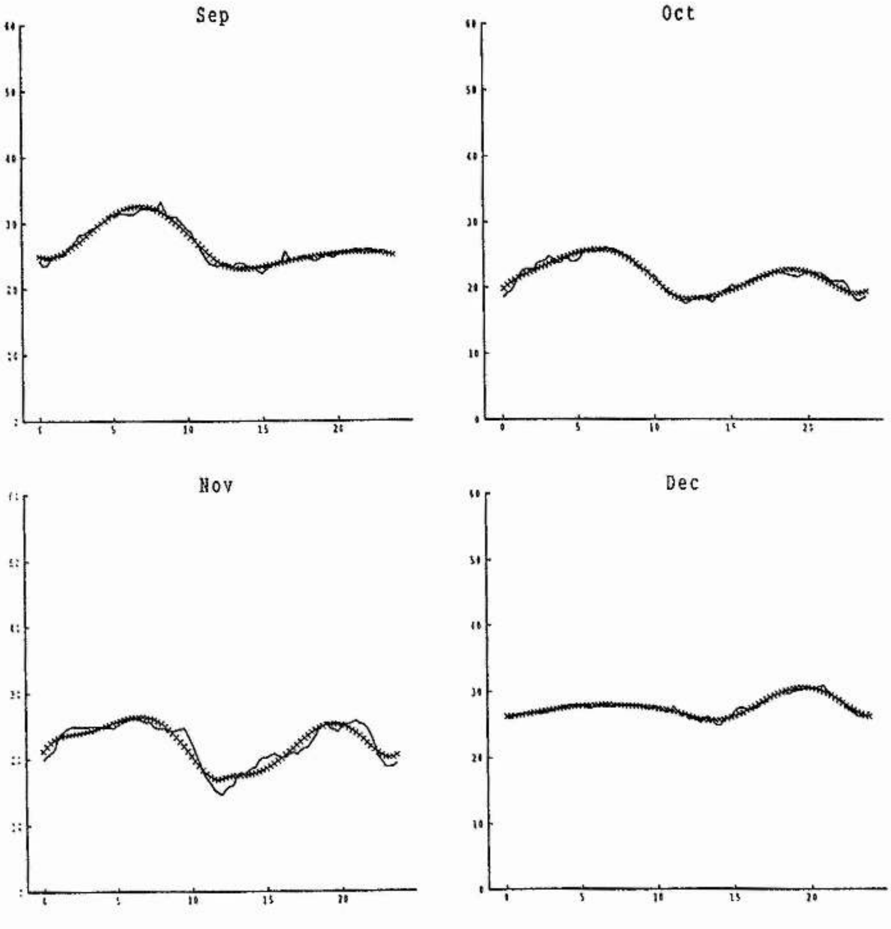


Table 4.1: Lack of fit (area between curves) for 6 modelling approaches

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Average
96 Par (Fig 4.5)	42.1	27.1	36.1	71.4	50.9	39.5	37.8	26.5	42.0	42.7	91.4	24.9	44.4
2 Corr PC (Fig 4.9)	89.2	216.5	60.5	243.3	935.2	379.1	107.3	223.3	183.6	234.2	135.2	217.3	252.0
3 Corr PC (Fig 4.11)	102.0	226.7	55.7	139.7	749.9	535.1	125.1	139.8	97.0	92.0	178.3	43.8	207.2
2 Cov PC (Fig 4.14)	2407	2541	2518	2299	3502	2287	2689	2535	2367	2221	2599	2784	2546
4 Cov PC (Fig 4.16)	2640	2591	2717	2636	3639	2469	2870	2701	2466	2397	2825	2921	2739
40 Par (Fig 4.19)	78.0	45.9	51.4	79.0	58.2	68.8	79.1	61.8	58.2	80.7	153.5	54.8	72.4

4.4: Parameter Reduction

Fig 4.6 shows the estimates of each of the eight parameters of the 2-quintic model over the twelve months of 1988. The parameters exhibit only 2 basic patterns: θ_6 and θ_8 show the same pattern, θ_5 and θ_7 show the inverse pattern to θ_2, θ_3 and θ_4, and θ_1 shows a similar pattern to θ_5 and θ_7. Because of this it may be possible to reduce the number of parameters in the model from eight to two or three for each month.

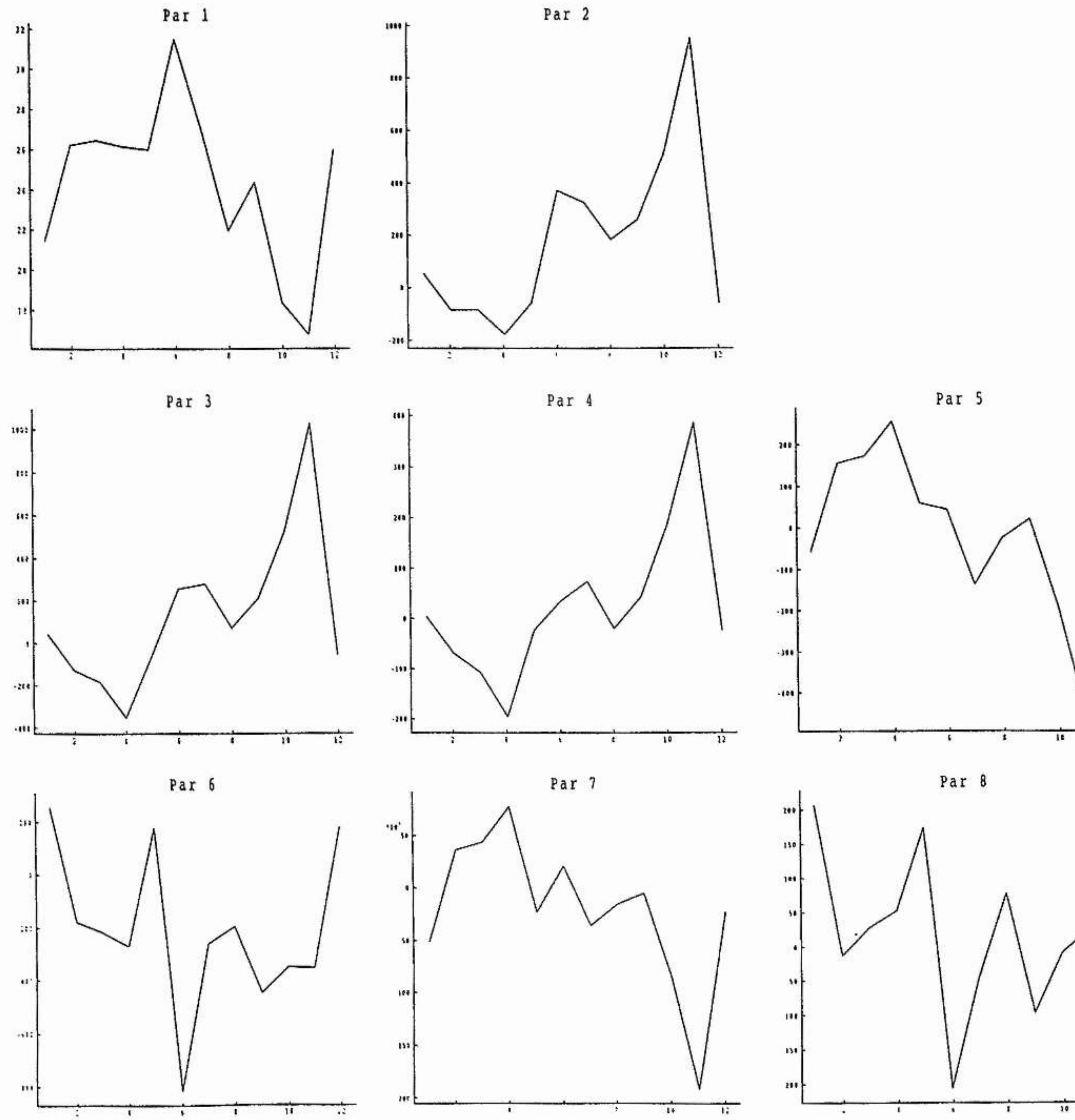
Using two parameters per month, say ψ_i and ξ_i (i=1...12), estimates for $\theta_1,...,\theta_8$ can be calculated as linear combinations of these new parameters by using a function of the form:

$$\theta_{ki} = \alpha_k\psi_i + \beta_k\xi_i \qquad\qquad (k=1...8)$$

This will reduce the number of parameters required to model a year from 96 to 40. Estimates of ψ_i and ξ_i will be required for each of the twelve months (total 24 parameters), and estimates of the eight α_k and β_k (total 16) will also be required.

Two possible approaches for obtaining estimates of ψ_i, ξ_i, α_k and β_k will be considered. The first method will be to use a principal components procedure on the original estimates of $\theta_1,...,\theta_8$, and the second will be based on reducing the dimensionality of the model directly and going back to the data.

Fig 4.6: Estimates of 8 parameters in the 2-Quintic model



First Method

To obtain estimates of ψ_i, ξ_i, α_k and β_k from a principal components analysis of the estimates of the θ_{ki} the following approach was adopted:

The principal components analysis gives the value of each principal component as a linear combination of the eight parameters:

$$PC_{mi} = h_m(\theta_{1i},...,\theta_{8i}) \qquad m=1...8 \ , \ i=1...12$$

(where PC_{mi} is the value of principal component m in month i)

If we set the last 6 principal components equal to zero (say), we obtain a set of simultaneous equations which can be solved to obtain each of the θ_{ki} as functions of the values of the first 2 principal components.

Thus:

set $PC_{mi} = 0$ $\qquad$ m=3...8 (say) , i=1...12

gives the set of equations:

$$PC_{1i} = h_1(\theta_{1i},...,\theta_{8i})$$

$$PC_{2i} = h_2(\theta_{1i},...,\theta_{8i})$$

$$0 = h_m(\theta_{1i},...,\theta_{8i}) \qquad m=3...8$$

which can be solved simultaneously to give the result:

$$\theta_{ki} = G_k\,(PC_{1i}\ , PC_{2i}) \qquad k=1...8 \ , \ i=1...12$$

This result allows calculation of estimates for the eight model parameters from the values of PC_1 and PC_2 obtained from the principal components analysis. By changing the number of PC's set to zero at the start of this procedure, it is possible to obtain estimates as functions of any number of the principal component scores, allowing flexibility in the number of parameters used for each month.

To this end, principal components analyses using both the correlation and covariance matrices were carried out on the parameter estimates. Both types of principal components analyses have been carried out as there is no obvious justification for being interested primarily in the shape or level of the parameters. This is because the primary aim of this procedure is not to model the parameters, but to obtain estimates of the parameters which will then be used to model the data. Tables 2 and 3 show the percentage variation amongst the parameters explained by the eight principal components obtained from both these analyses.

Table 4.2: Correlation PC

PC	1	2	3	4	5	6	7	8
%var	66.87	28.51	3.77	0.52	0.32	0.01	0	0

Table 4.3: Covariance PC

PC	1	2	3	4	5	6	7	8
%var	83.55	16.14	0.18	0.14	0.00	0.00	0	0

Considering the correlation analysis first, the first two PC's account for 95.38% of the observed variation amongst the parameters. If the third PC is included, this value increases to 99.15%. PC's 4,...,8 make much smaller contributions to the percentage variation explained and are not going to be considered.

Fig 4.7 shows that the first two correlation PCs follow the two basic patterns observed in the parameters. Fig 4.8 allows comparison of the estimates for $\theta_1,...,\theta_8$ calculated using the above principal components method and their original estimates. The original parameter estimates are closely approximated. θ_1 is the parameter least well approximated. As $f_1=1$, θ_1 represents the actual ozone level at x=0 , or 7pm, and is added to the predicted ozone values at all other times. This will lead to an error in the level taken by the model.

Fig 4.9 shows that the ozone curves generated by this approach do not fit the data well when compared to the original model, especially for the months of May and June. This is confirmed by the lack of fit results displayed in table 4.1. The average lack of fit for this approach is over 5 times greater than that for the original model. There is a slight error in the level of the model values for many of the months, due to the error in θ_1, and the model has failed to reproduce the correct shapes for many of the monthly curves.

To investigate if the lack of fit can be improved by including the third correlation principal component, estimates of $\theta_1,...,\theta_8$ were calculated as functions of the first three correlation principal component scores. Fig 4.10 allows comparison of these estimates and the originals. This method provides more accurate estimates of $\theta_1,...,\theta_8$, especially $\theta_{1.}$ Fig 4.11 shows that despite this improvement in the estimates of the parameters the inclusion of the third principal component has not greatly improved the fit of the model to the data. This is confirmed by the lack of fit results displayed in table 1, there has been a slight drop in the average lack of fit, but it is very small compared to the difference between these two correlation PC approaches and the original approach. There has been a slight improvement in the fit for some months but the noted discrepancies for the months of May and June still exist.

Moving on to the covariance analysis, the first two PC's account for 99.69% of the observed variation amongst the parameters. If a larger number of PC's are to be considered

Fig 4.7: Values of correlation PCs

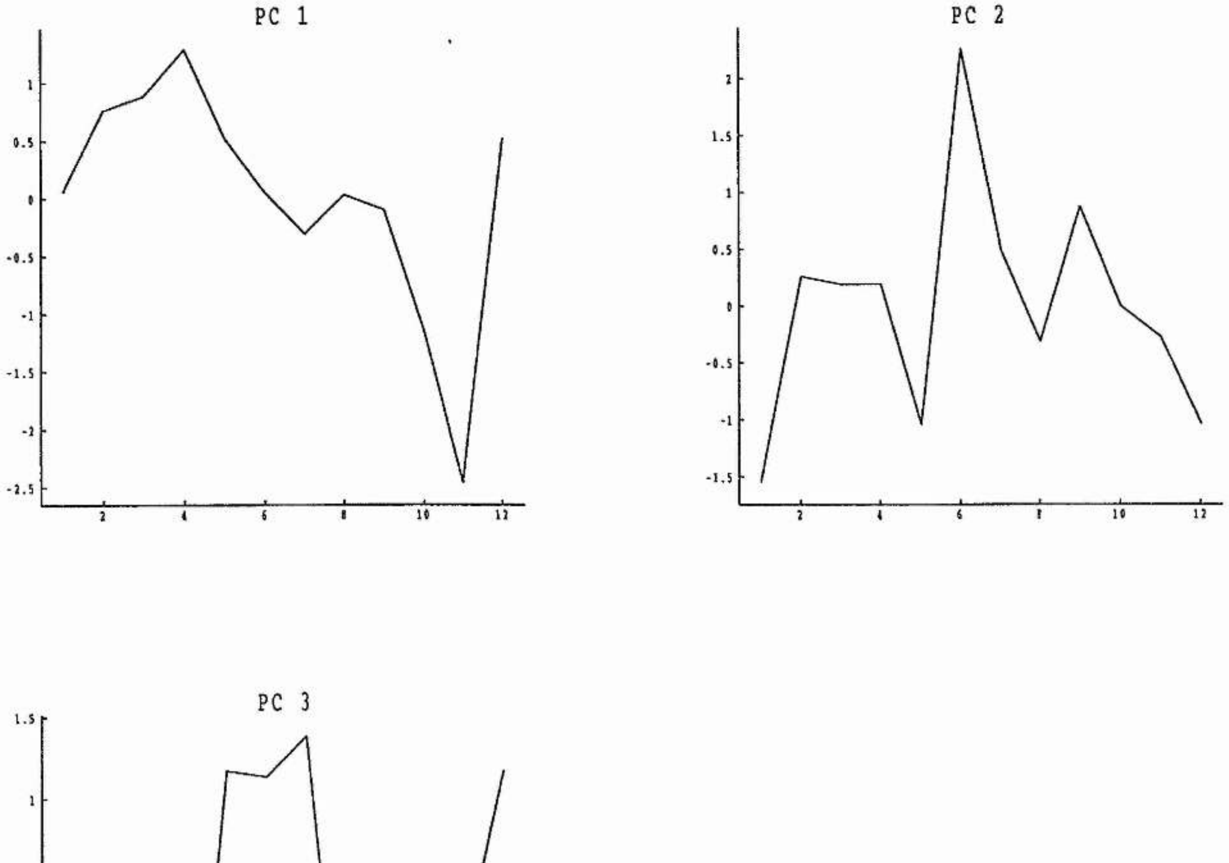


Fig 4.8: Estimates of θ_1-θ_8 as f(PC1,PC2)

—— original estimates

----- f(PC1,PC2)

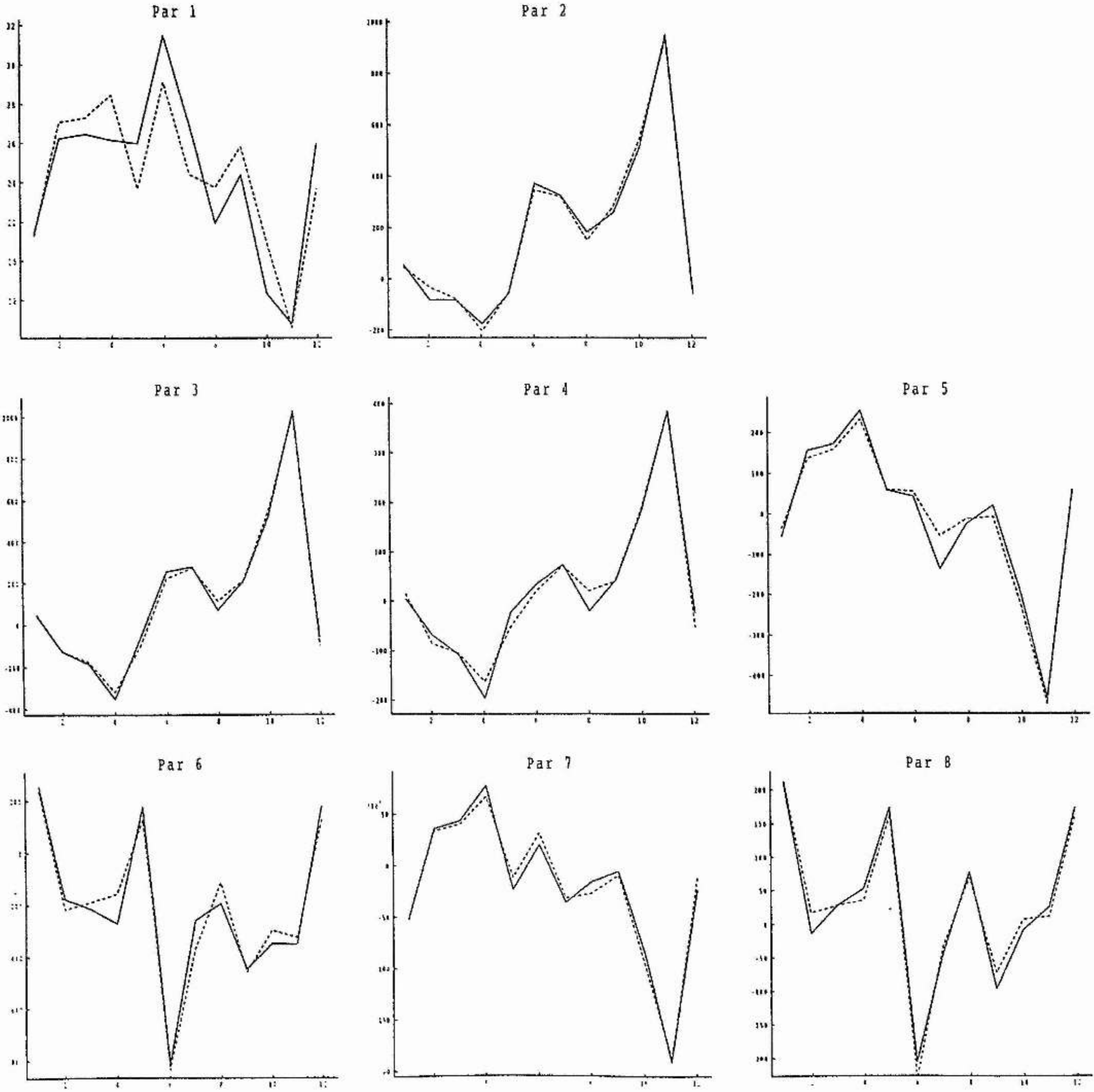


Fig 4.9(part 1): Fit by first 2 correlation PC approach

—— raw data
xxxx model fit

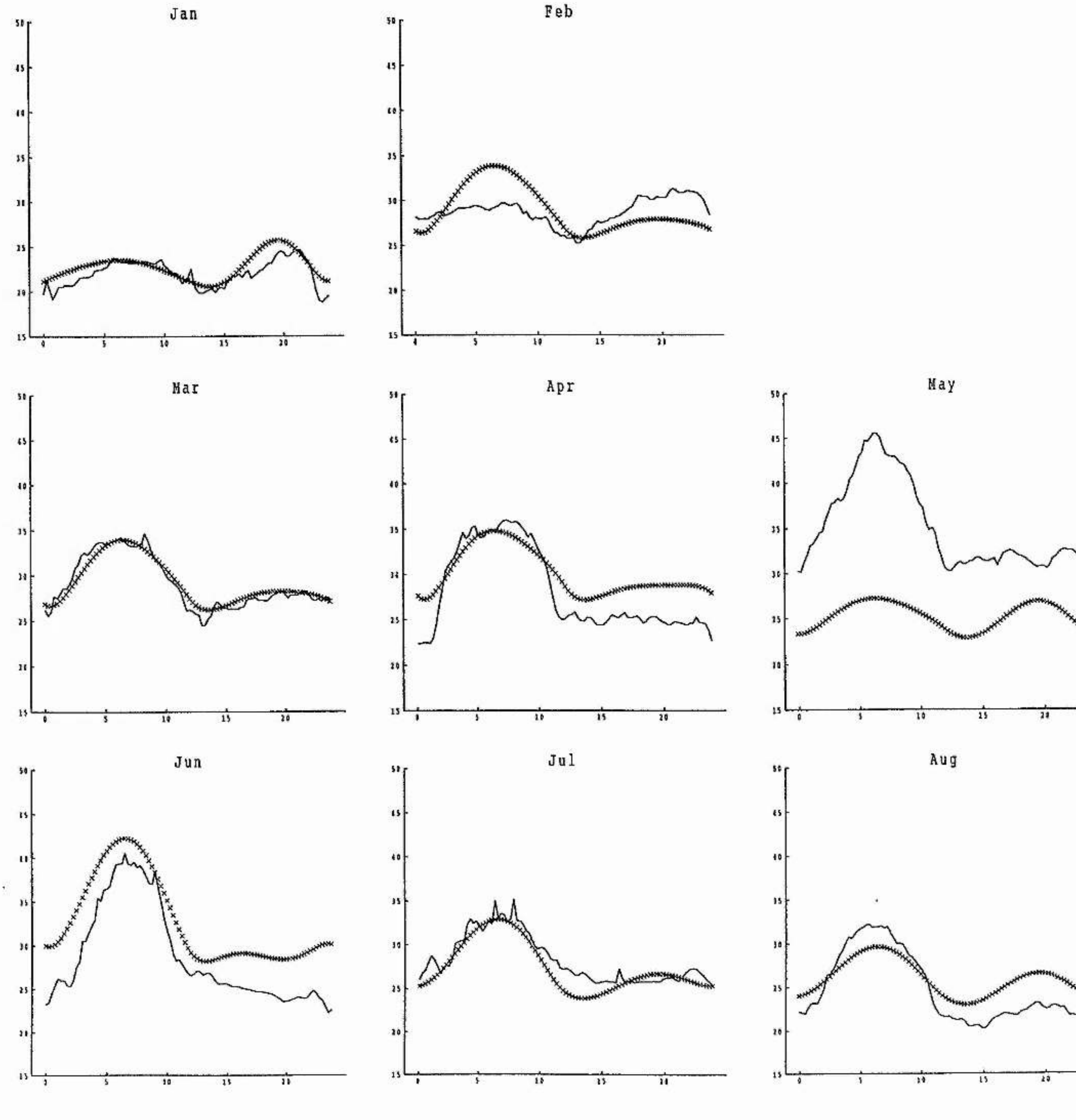


Fig 4.9(part 2): Fit by first 2 correlation PC approach

——	raw data
xxxx	model fit

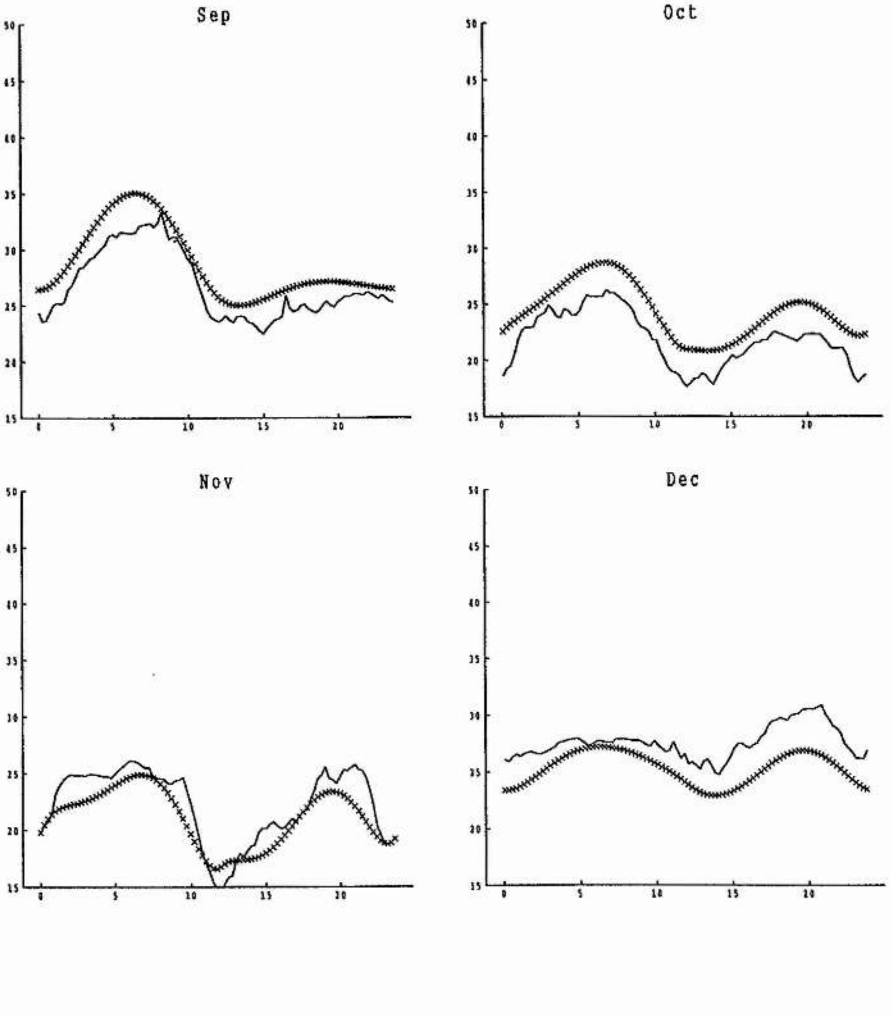


Fig 4.10: Estimates of θ_1-θ_8 as f(PC1,PC2,PC3)

—— original estimates
----- f(PC1,PC2,PC3)

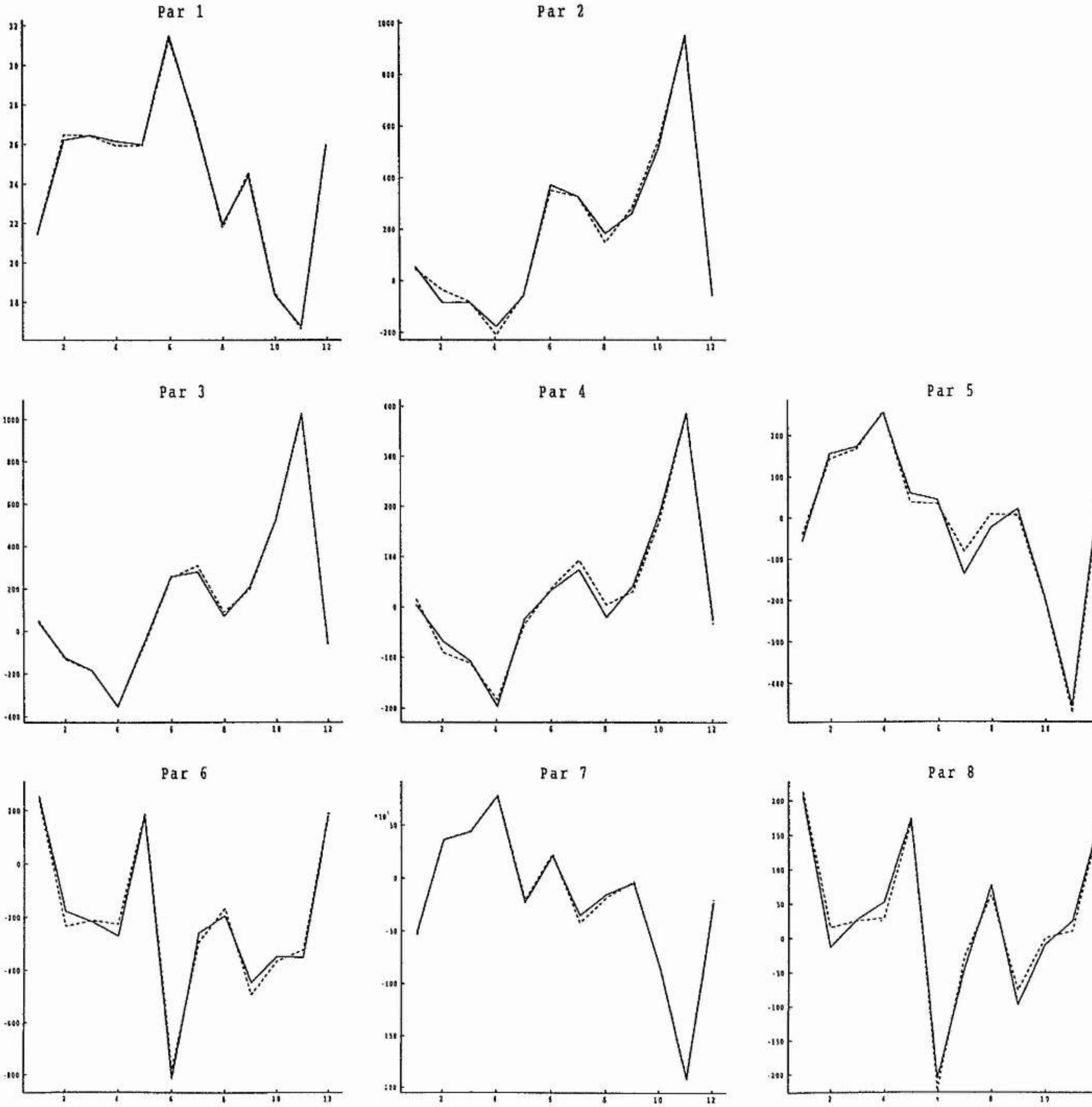


Fig 4.11 (part 1): Fit by first 3 correlation PC approach

—— raw data
xxxx model fit

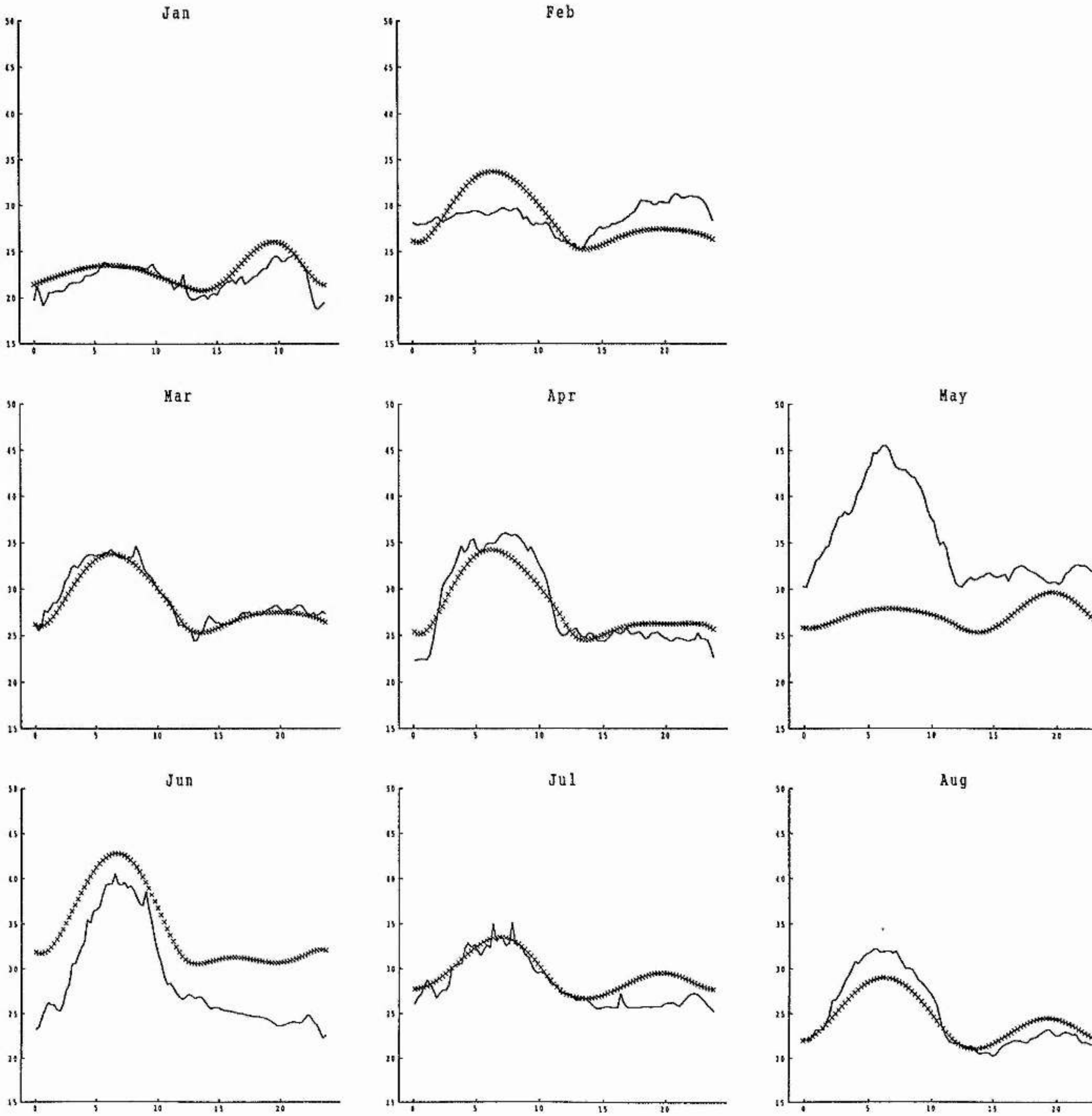
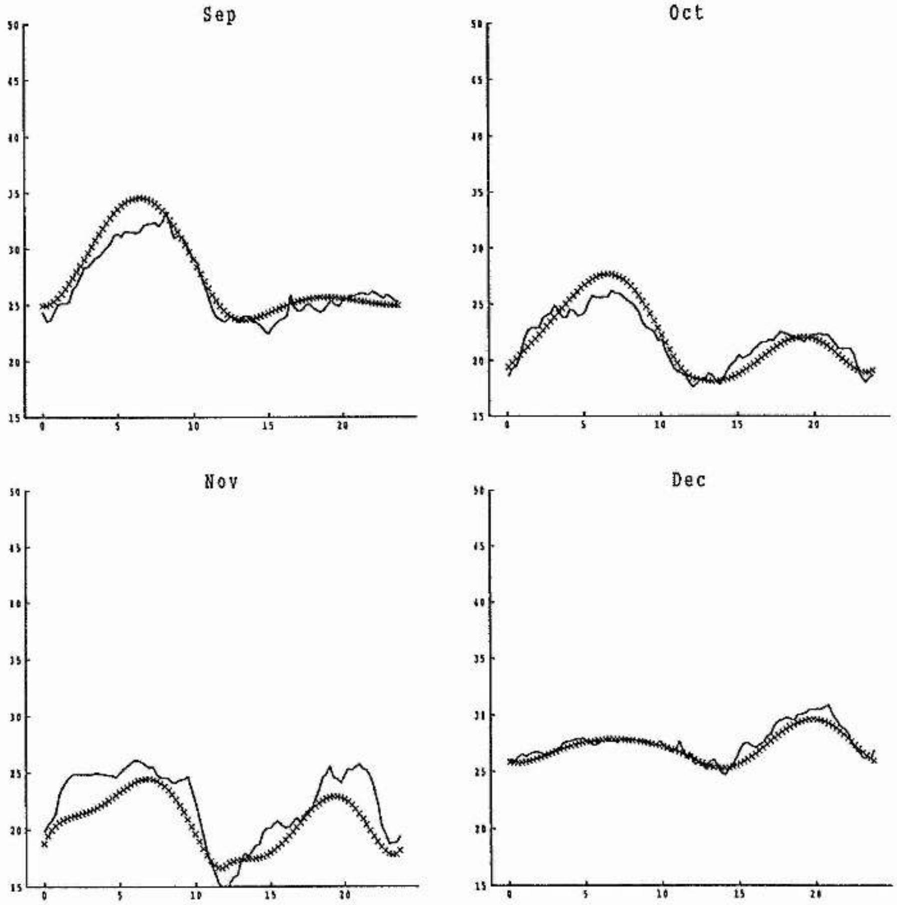


Fig 4.11 (part 2): Fit by first 3 correlation PC approach

— raw data
xxxx model fit



it is necessary to include the third and fourth PC's as they both make a very similar contribution to the percentage variance explained and the contribution made by PC's 5,...,8 are zero to 2 decimal places. If the third and fourth PC's are included, the percentage variance explained increases to almost 100%.

Fig 4.12 shows that the first two covariance PCs also follow the two basic patterns observed in $\theta_1,...,\theta_8$. Fig 4.13 allows comparison of the estimates for $\theta_1,...,\theta_8$ calculated using the first two covariance PC's and the original estimates. The original parameter estimates are less well approximated than by both the correlation PC methods. The discrepancy for θ_1 has now been greatly increased, as the use of the covariance matrix has reduced its importance to the analysis due to its small magnitude relative to the other seven parameters. θ_7 is now very well approximated, as it is by far the greatest in magnitude and as such has the highest importance to a covariance PC analysis. Fig 4.14 shows that the fit of the model to the data is extremely poor, there being now an error of approximately 30 ppb in the level of the curves generated by the model. As discussed above this is a result of a similar error in the level of the estimates for θ_1. This is reflected in the lack of fit results in table 4.1, the average now being an order of magnitude greater than the two correlation PC approaches.

Fig 4.15 shows estimates for $\theta_1,...,\theta_8$ calculated as functions of the first four covariance principal component scores. These estimates are closer to the originals, except for θ_1, which still exhibits a large error in level. Fig 4.16 confirms that this error in the level of θ_1 leads to a similar error in the level of the model fit, although except for the month of May, the shape of the fit is good. The average lack of fit is now even greater than the two covariance PC approach.

The principal components approach considered above has been shown not to reproduce the monthly curves well. The reason is that in a principal components analysis the focus is the variation amongst the estimated parameters, not the fit of the model to the data. The 'data'

Fig 4.12: Values of covariance PCs

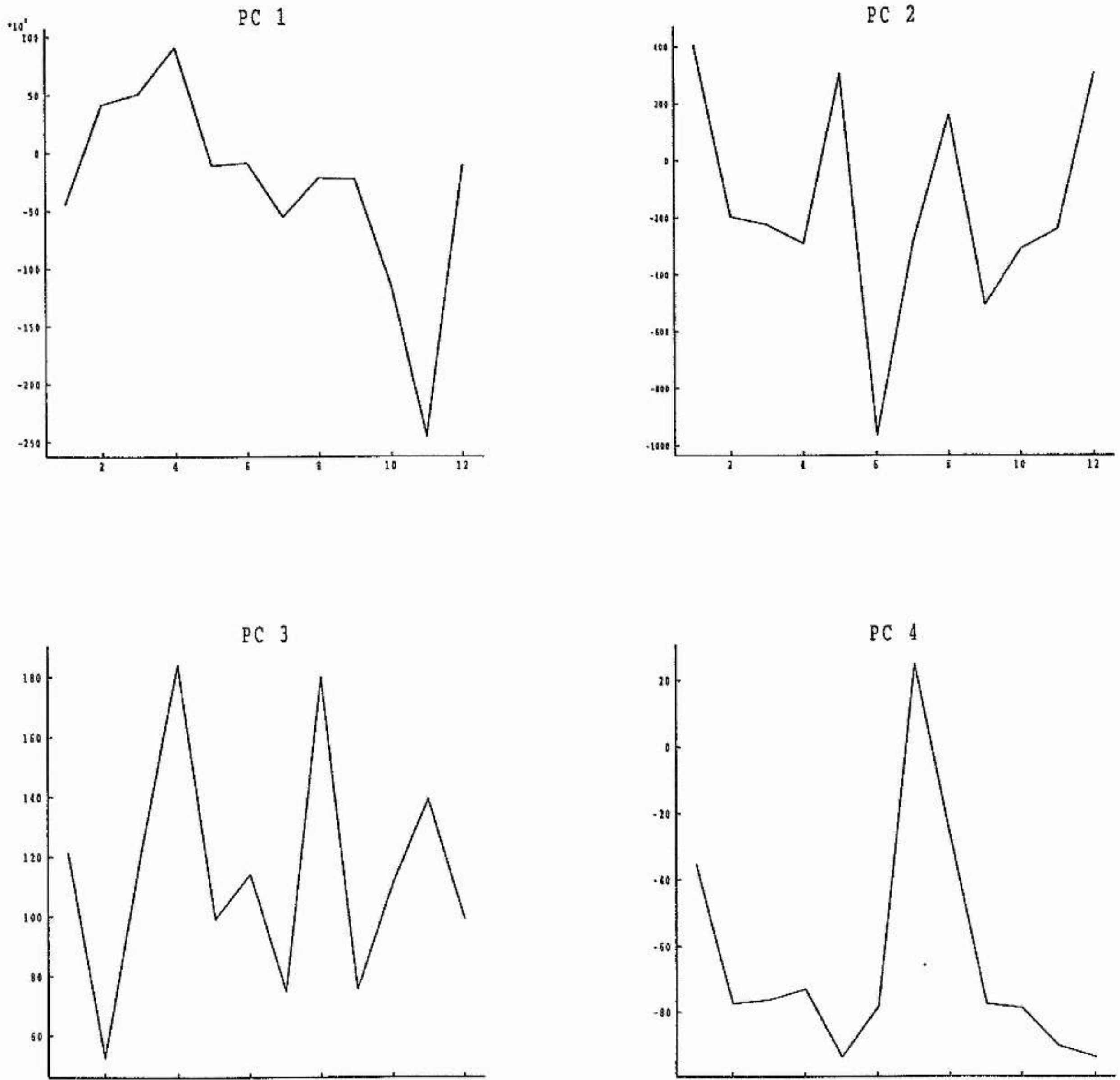


Fig 4.13: Estimates of θ_1-θ_8 as f(PC1,PC2)

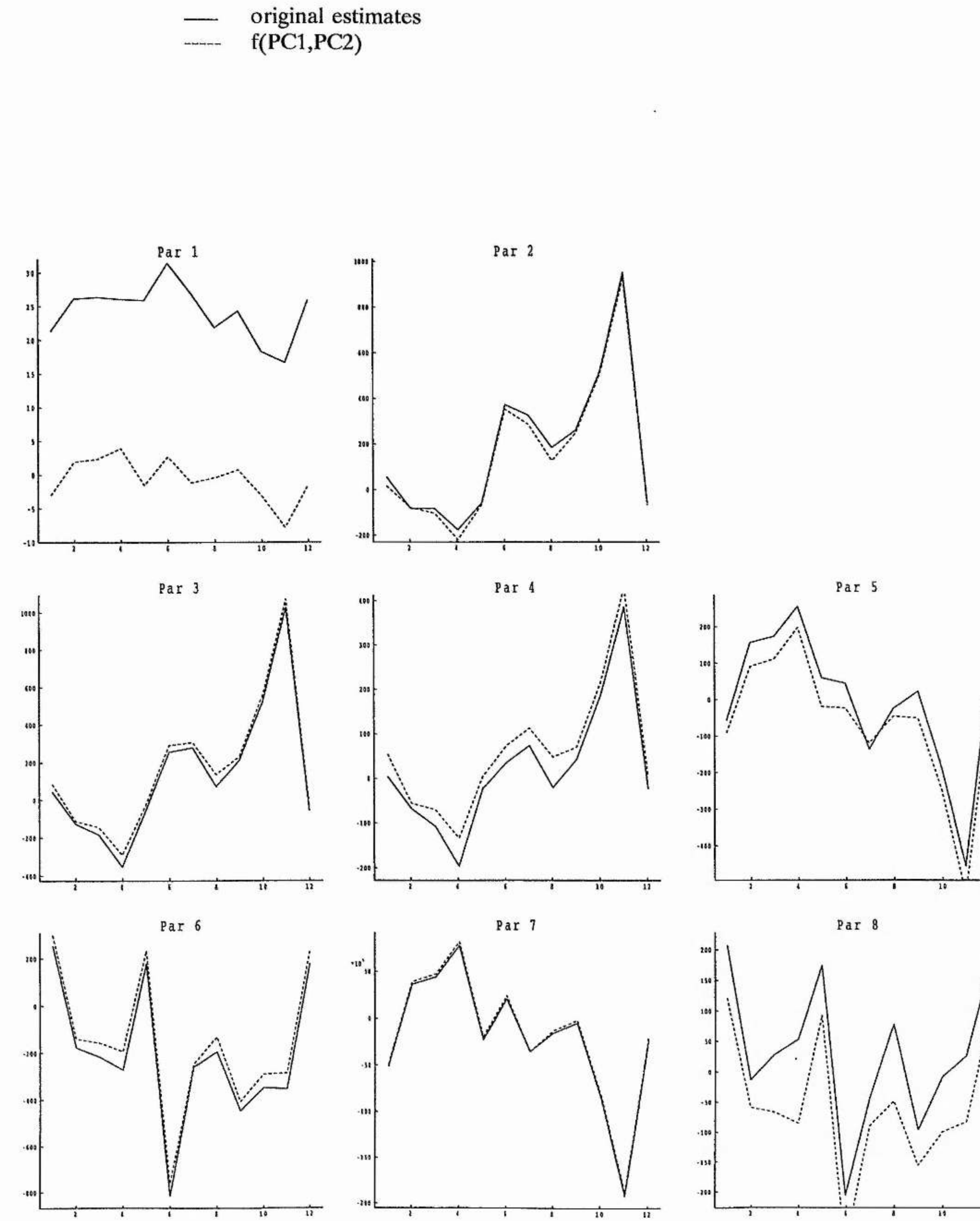


Fig 4.14 (part 1): Fit by first 2 covariance PC approach

—— raw data
xxxx model fit

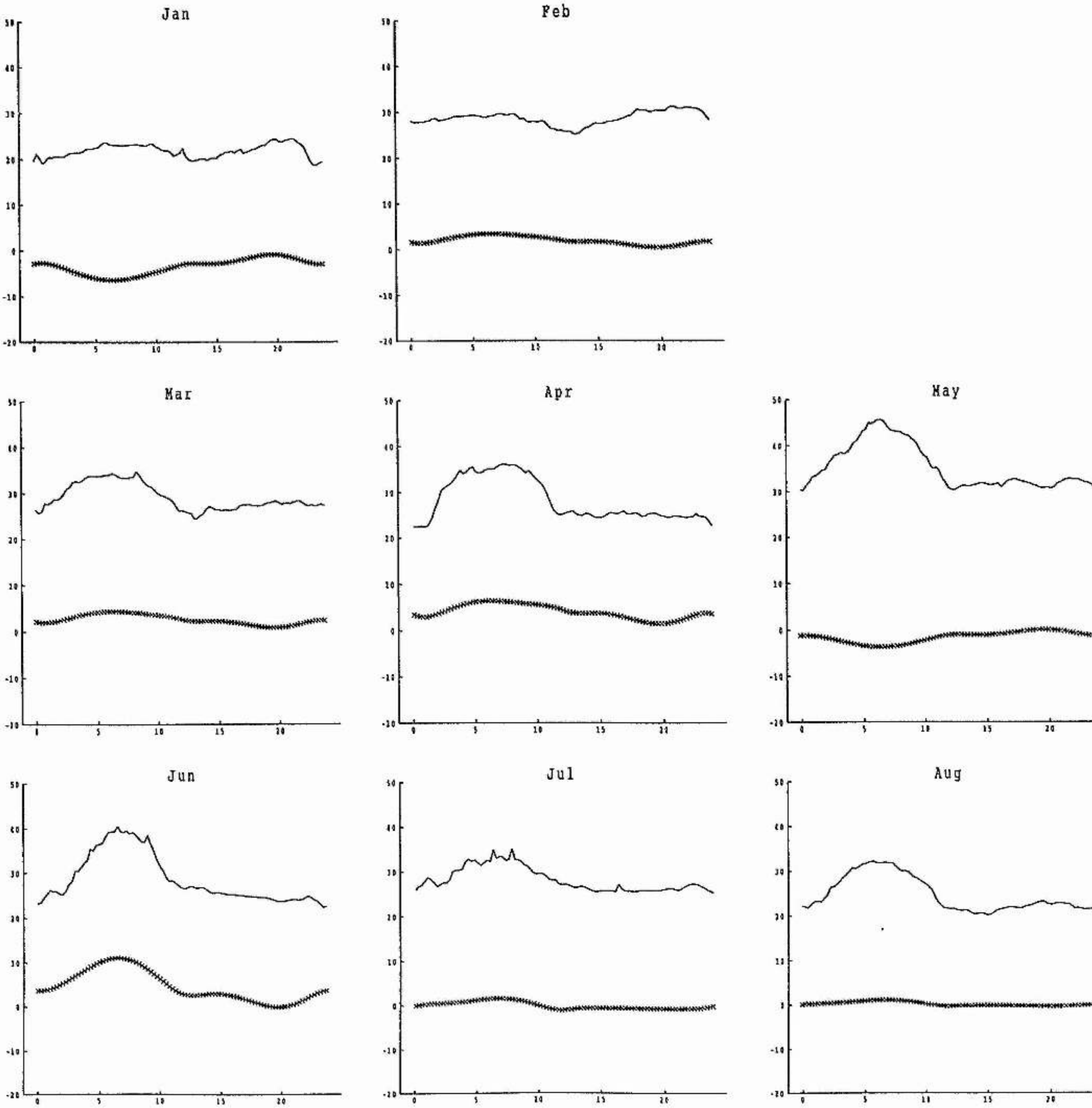


Fig 4.14 (part 2): Fit by first 2 covariance PC approach

—— raw data
xxxx model fit

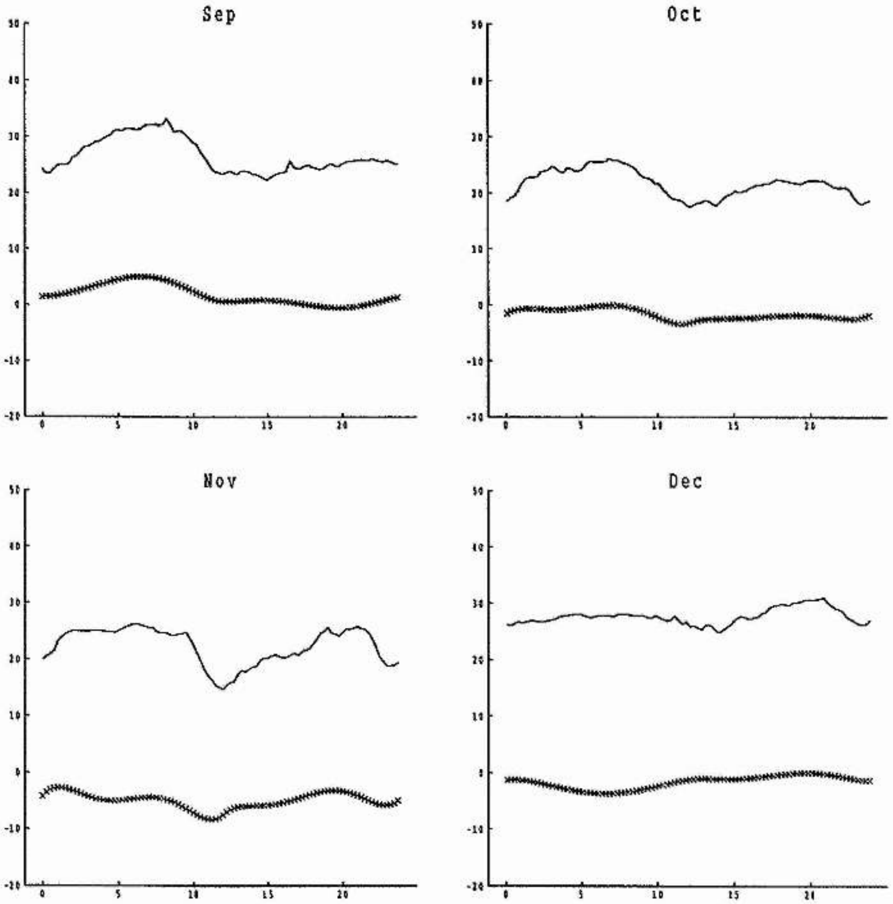


Fig 4.15: Estimates of θ_1-θ_8 as f(PC1,PC2,PC3,PC4)

—— original estimates
----- f(PC1,PC2,PC3,PC4)

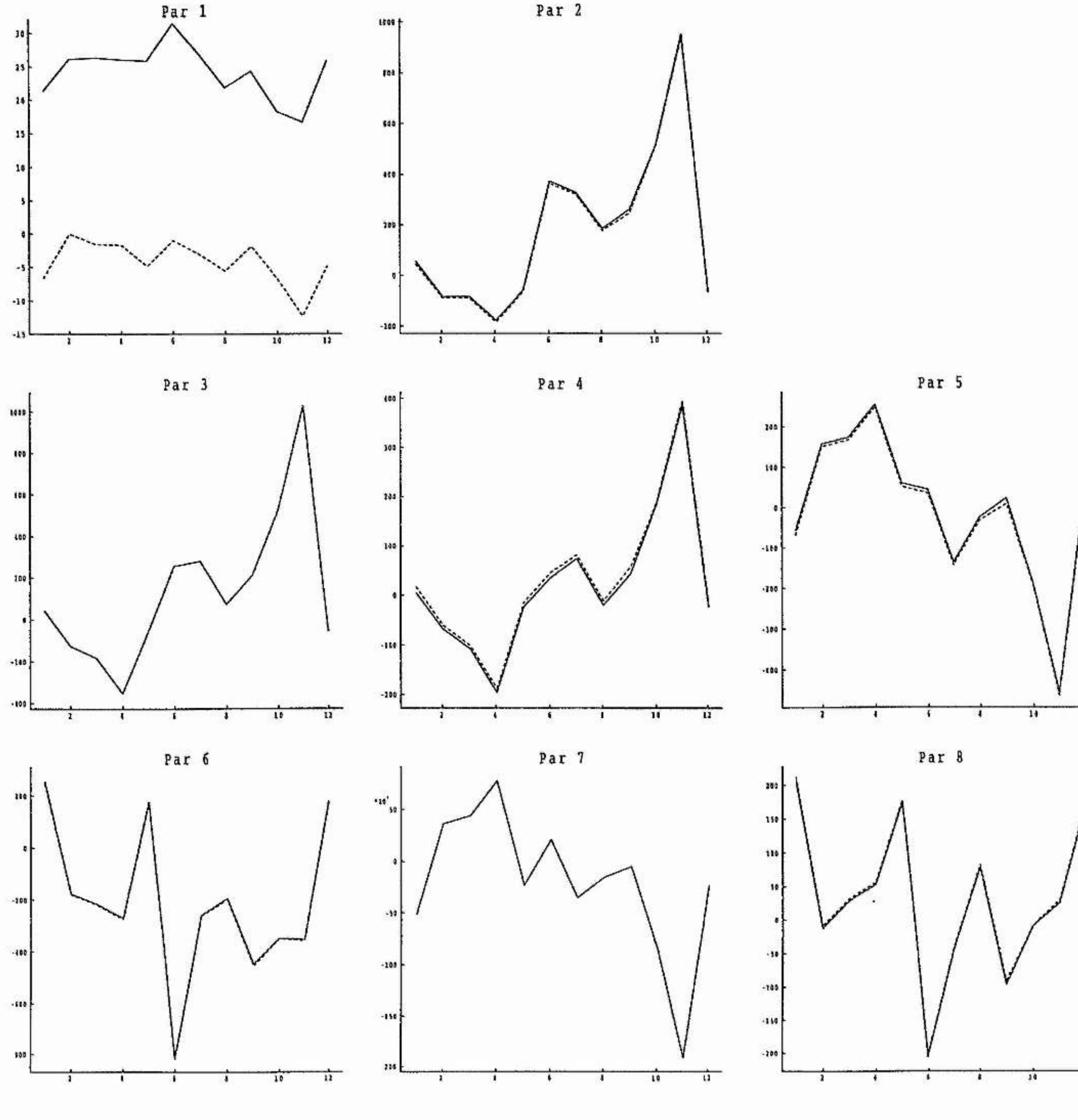


Fig 4.16 (part 1): Fit by first 4 covariance PC approach

—— raw data
xxxx model fit

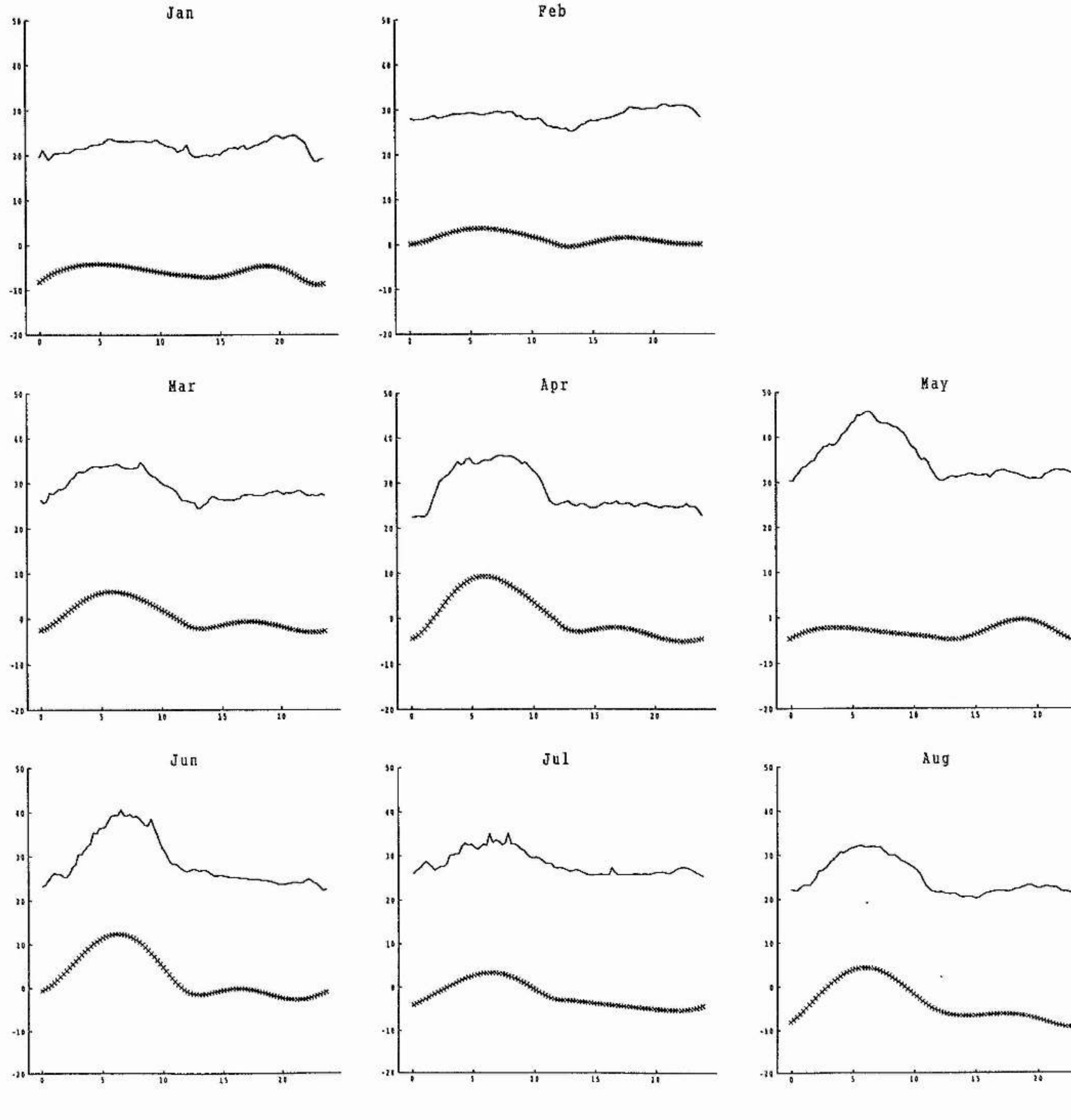
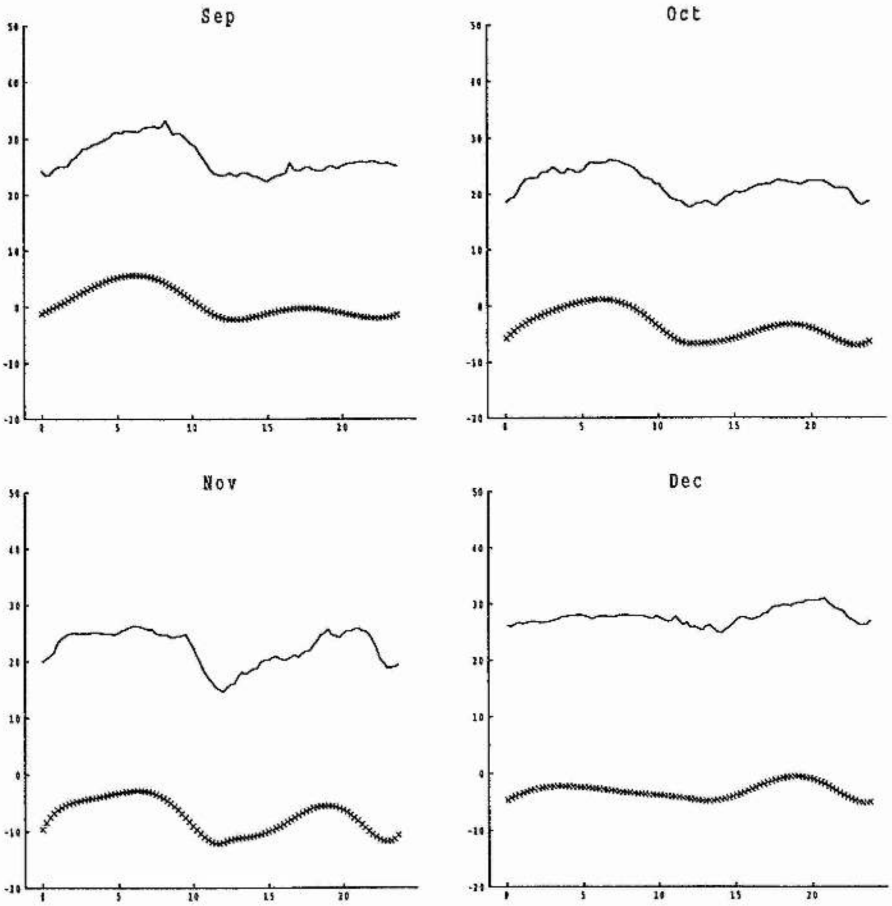


Fig 4.16 (part 2): Fit by first 4 covariance PC approach

—— raw data
xxxx model fit



for the PC analysis are the estimated parameters from another model. The results displayed in figs 4.8,4.10, 4.13 and 4.15 indicate that some of these methods have been reasonably successful in achieving an approximation to the parameters, but that these approximations do not translate into a good model fit to the data. In addition, this approach also has the disadvantage that the original model must first be fitted to the data before the parameter reduction is carried out.

Second Method

An alternative method is to reduce the dimensionality of the original 2-quintic model directly. In

$$E[y_{ij}] = \sum_{k=1}^{8} \theta_{ki}\, f_{kj} \qquad\qquad i=1...12\ ,\ j=1...96\ ,\ k=1...8$$

write each θ_{ki} as a linear combination of two new parameters, ψ_i and ξ_i, as follows:

$$E[y_{ij}] \;=\; \sum_{k=1}^{8} (\alpha_k\psi_i + \beta_k\xi_i)\, f_{kj}$$

$$= \;\psi_i \sum_{k=1}^{8} \alpha_k f_{kj} \;+\; \xi_i \sum_{k=1}^{8} \beta_k f_{kj}$$

This approach can be related to the idea of generating two patterns, and modelling the monthly curves as weighted averages of these. These two shapes (S_1 and S_2, say) are generated by considering the:

$$S_1 = \sum_{k=1}^{8} \alpha_k f_{kj} \qquad \text{and} \qquad S_2 = \sum_{k=1}^{8} \beta_k f_{kj}$$

terms in the model. Thus the 16 α_k and β_k give the two shapes, and the 24 ψ_i and ξ_i allow mixing of them in various proportions to model the data.

The expected advantage of writing the model in this way is that the suggested reduction in parameters has occurred, but the reduced parameter model is now being fitted to the data directly. This model can be fitted iteratively by a least squares procedure: given values for (ψ_i , ξ_i) then estimates of (α_k , β_k) are obtainable by inversion of a 16x16 matrix; given values for (α_k , β_k) then estimates of (ψ_i , ξ_i) are obtainable separately for each i by inversion of a 2x2 matrix. The procedure is thus to iterate until the values of all the parameters have stabilised.

A FORTRAN program has been written to implement this approach to the data (Listing in Appendix 1). The user must enter initial estimates of the parameters, and the program will then iterate until the relative change in all parameters is less than 0.01% between iterations. A range of randomly generated initial values in the range [1,20] were tried and the final parameter estimates were found to be insensitive to these initial values.

Fig 4.17 shows the estimates of the two shapes S_1 and S_2 introduced above plotted together for comparison. They do not have any apparent physical interpretation. They are very similar in shape, having maxima and minima at the same time, but differ primarily in level. S_2 also has a slightly smaller spread. An examination of the estimates of ψ_i and ξ_i reveals that S_1 and S_2 are mixed in the ratios -1 : A where $A \in (1.01, 1.06)$. Fig 4.18 shows combinations of $\psi_i S_1$ and $\xi_i S_2$ for various values of ψ_i and ξ_i. The shape of the curve is determined by the ratio $\psi_i : \xi_i$. As the value of A
$(-\xi_i / \psi_i)$ is increased from 1.01 to 1.06, the relative sizes of the daytime and night time 'humps' vary. The level is controlled by varying the magnitude of ψ_i and ξ_i, for example in fig 4.18 -1:1 and -16:16 have exactly the same shape, but different levels.

Fig 4.19 shows the ozone curves generated by this approach. This reparameterisation generates curves very close to the original model and the data, but with 40 parameters for a year instead of the original 96. The lack of fit results are displayed in table 4.1. The average lack of fit is slightly higher than the original model, but is considerably lower than all four PC approaches. Thus the parameter reduction has been successfully achieved, without greatly sacrificing the fit of the model to the data.

Fig 4.17: Estimates of S_1 and S_2

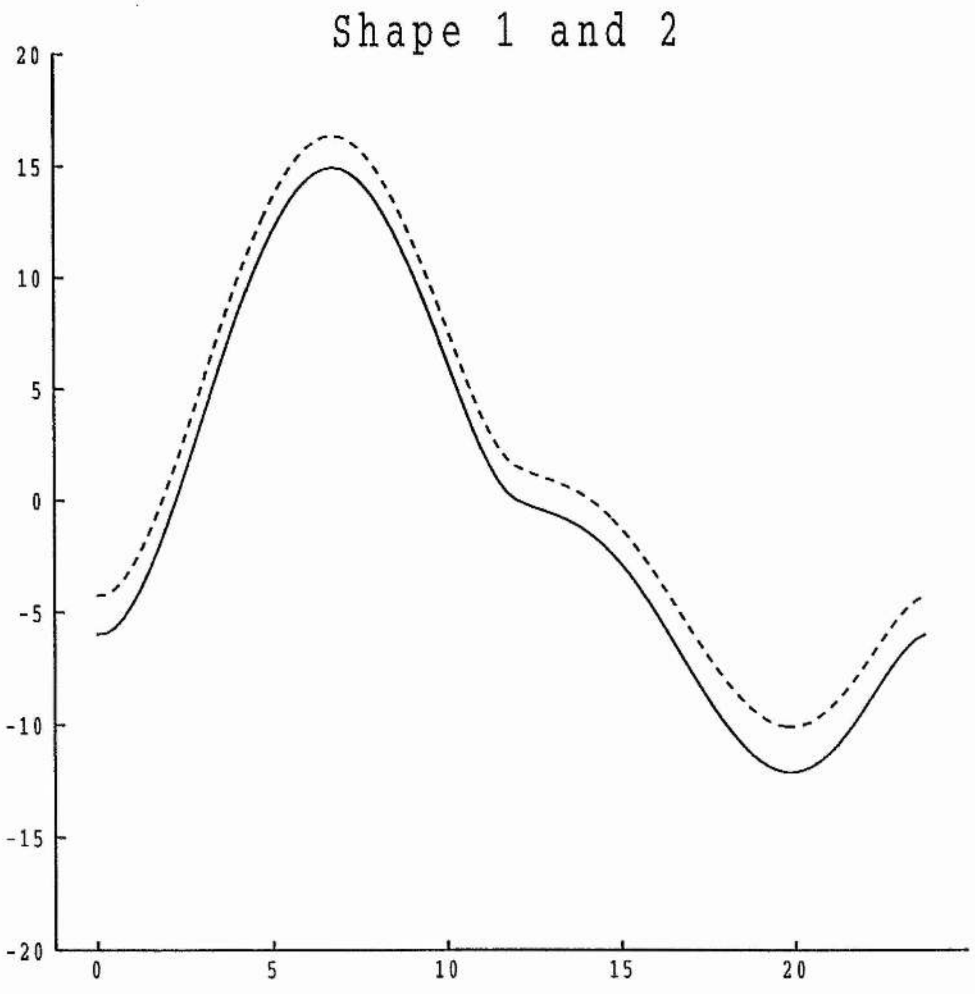
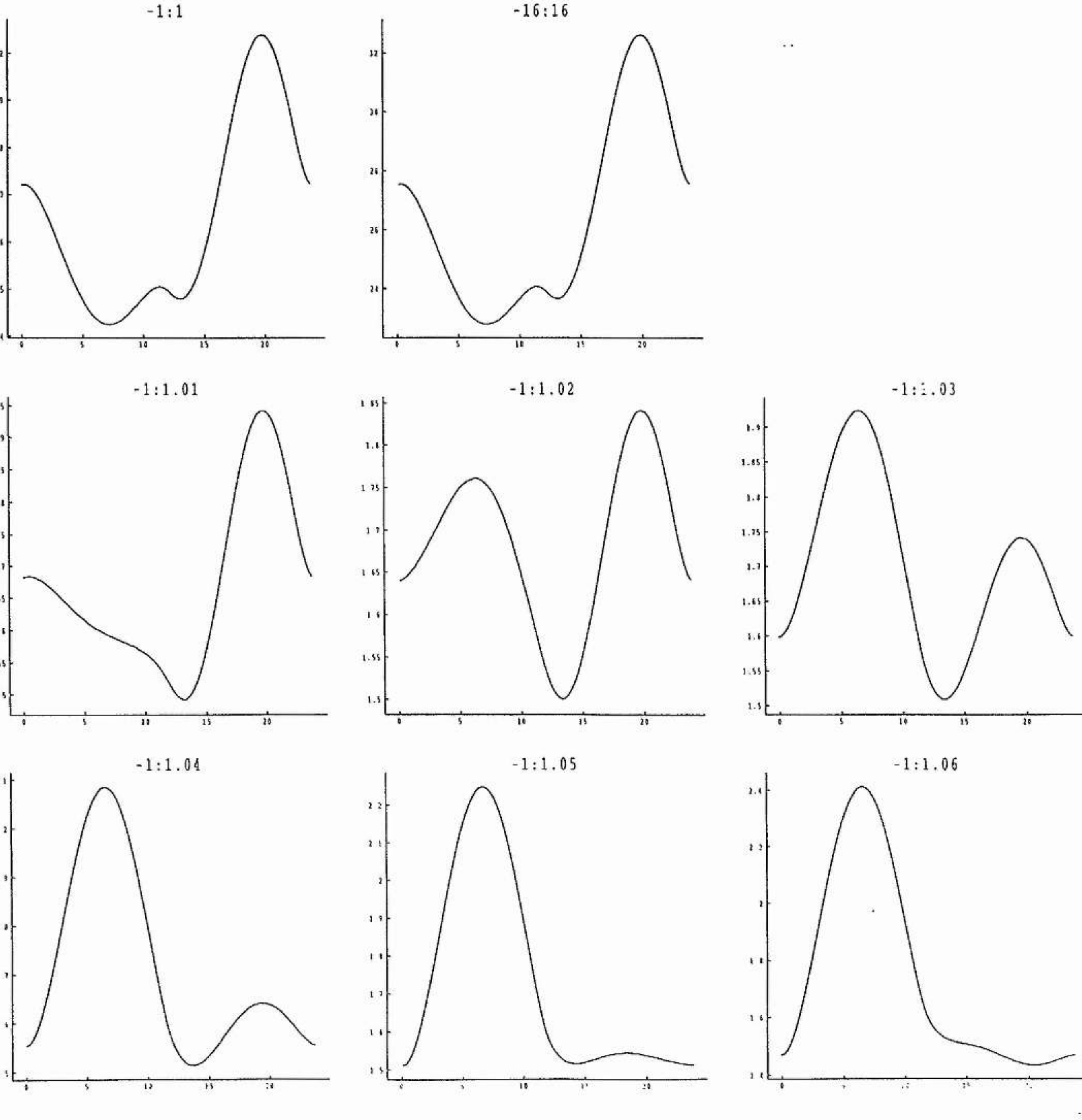


Fig 4.18: Combinations of $\psi_i S_1$ and $\xi_i S_2$


-1:1

-16:16

-1:1.01

-1:1.02

-1:1.03

-1:1.04

-1:1.05

-1:1.06

Fig 4.19 (part 1): Model fit using reparameterisation

— raw data
xxxx model fit

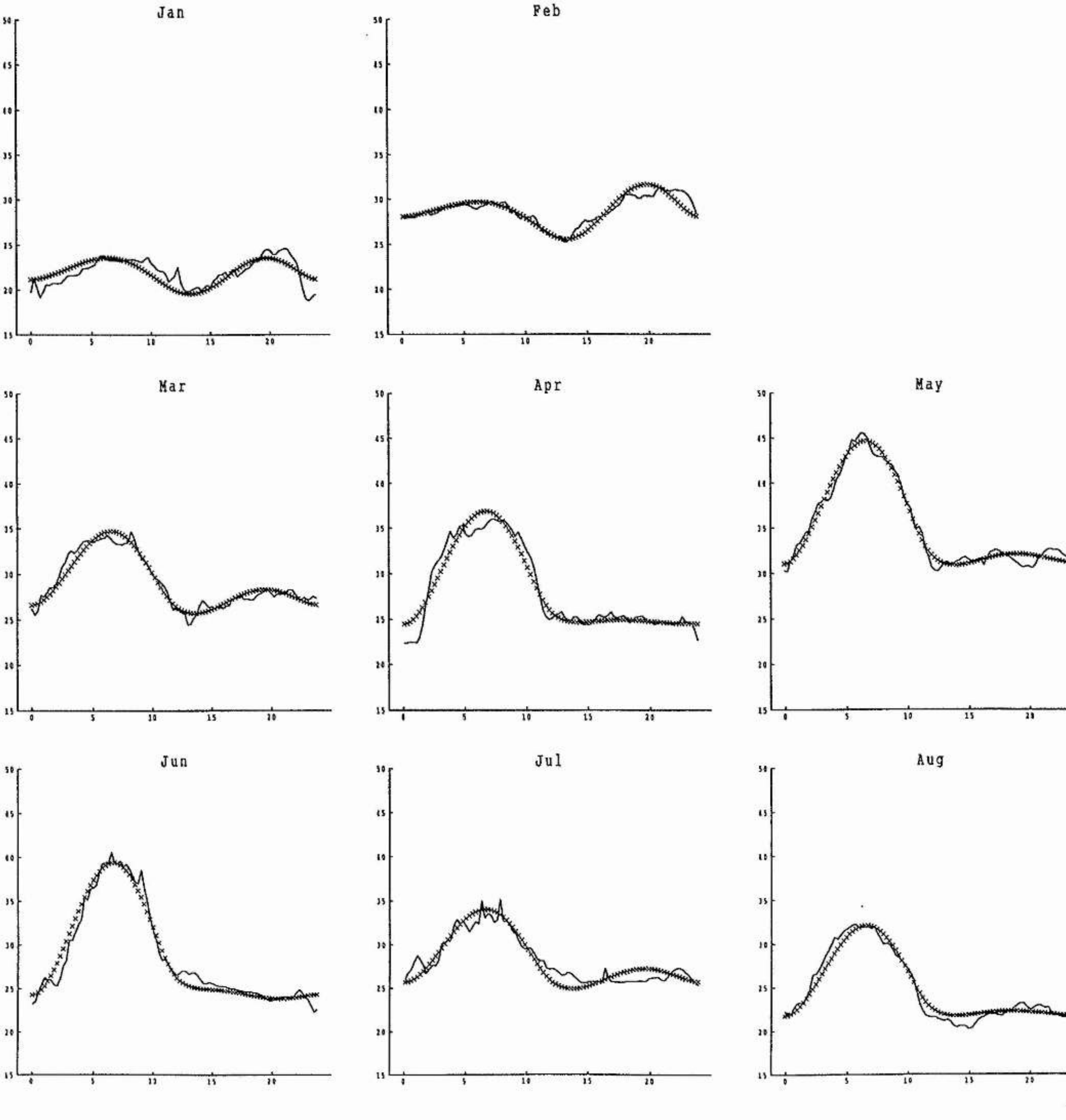
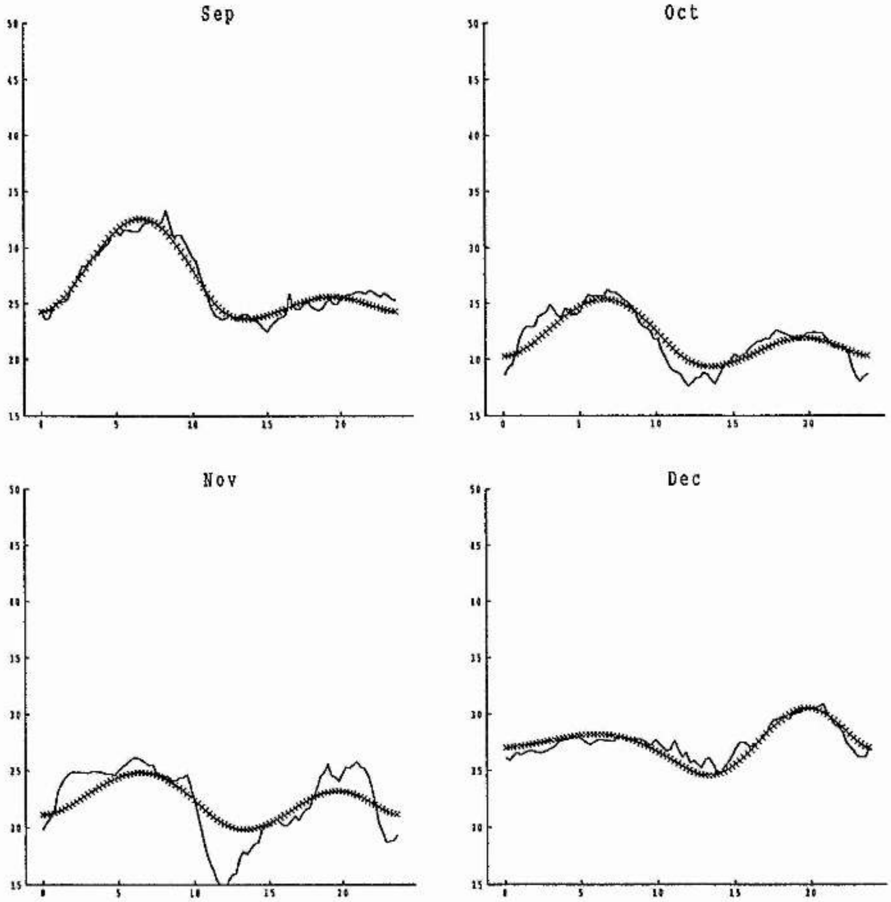


Fig 4.19 (part 2): Model fit using reparameterisation

—— raw data
xxxx model fit



4.5: Conclusion

The work carried out in this chapter has demonstrated that there is an observable variation in the shape, level and spread of the average ozone diurnal cycle from month to month over a year.

The use of the transformed time axis has shown that it is possible to explain much of the observed variation in shape by merely considering the position of the sun in the sky. Shifting the start of a day to 7am on this transformed time axis indicated that the average diurnal cycle can be split up into a daytime and a night time section, both of which can be considered separately.

The first model considered in this chapter has been shown to effectively model all twelve monthly curves (a total of 12x96=1152 observations), using 96 parameters. However, an examination of the estimates of these parameters indicated that it should be possible to greatly reduce the number of parameters in the model. None of the principal components methods considered were successful in achieving this reduction without sacrificing the fit of the model to the data, but the approach of reducing the number of parameters in the model directly was successful in modelling the data using a much reduced total of 40 parameters instead of 96.

Chapter 5: Classification of Individual Days

5.1: Introduction

The work carried out in the previous chapter has shown that the average diurnal ozone curve changes from month to month throughout the year. It is hypothesised that this change is due to the existence of several different types of day, whose relative frequencies change from month to month throughout the year. This would provide a possible explanation for the difference between the average curves observed for each month, with the monthly average curve being the appropriate weighted average of the typical diurnal curves.

It would be of great interest to find these types, and examine both their shapes and occurrences to provide possible physical causes. One might expect them to correlate to the weather pattern present during each day, since the rates of the various processes which account for the observed level of ozone depend on the weather. This chapter describes the use of cluster analyses to obtain estimates of these types.

The approach taken in this thesis has been to discover patterns in the ozone data without reference to known physical processes, and to seek any correspondence between these patterns and possible causes. Thus there is no use made of any data available on these processes (eg. Wind speed & direction, temperature, solar radiation etc.) until the patterns have been estimated by purely statistical means. The complementary approach, developing physically-based models for ozone occurrence directly, is being carried out by others.

5.2: Data

The raw data are the same as in the previous two chapters, 15 minute averages of measurements on levels of ground level ozone, measured in parts per billion (ppb), at ITE, Bush Estate, Midlothian during 1988, with all the known `spikes` replaced. To reduce the volume of data, these 15 minute readings have been aggregated to give hourly averages. This has the advantage of smoothing the data, reducing the effect of the outliers to which the 15 minute data set is prone.

As in the previous chapter, the time axis was transformed to make sunrise, sunset, midday and midnight occur at the same transformed time each day . Note that midday and midnight are taken to refer to the midpoints of day and night, not 12:00 and 00:00. Each day is represented by a vector ($\underline{y}_i$) containing 24 values, at equally spaced points on the transformed time axis, obtained by linear interpolation between the hourly values. As before, days are split into a `daytime` section and a `night time` section, each having 12 values. Because ozone damage to plants only occurs when the plant stomata are open, the daytime sections may be of particular interest.

One disadvantage of partitioning time into days is that the weather can change during a day. If the types of ozone curve uncovered do indeed relate to the weather during the day then the ozone curve would then begin as one type and end as another. Taking half a day as the object for a classification will reduce the occurrence of this problem.

In this data set, 30 out of the 366 days contain large sections of missing values. These days have been excluded from this analysis, leading to a data set consisting of 336 objects, each consisting of 24 hourly observations.

5.3: Methods

In this analysis an object is the series of 24 observations from the whole day or 12 observations from the day or night time sections. It is necessary to construct some measure of dissimilarity between two such objects to be used for whatever classification algorithm is chosen.

The level and spread of a day's ozone curve depends more on the behaviour of previous days than the weather during that day. Thus we are interested in attempting to classify the different types of day by considering the shape of their ozone curves. This will affect the choice of dissimilarity measure to be used.

The classical choice of dissimilarity measure in cases where one is more interested in a difference in shape would be to use the sample correlation coefficient calculated between two objects (Gordon 1987). One problem with this measure is that the correlation coefficient is invariant under permutation of the variable labels which in this case represent time ordering. As the data here have a definite time ordering it would be preferable in the interest of using as much of the information available as possible to use a measure which takes account of this over one which does not.

Another possibility for the choice of dissimilarity measure would be to use the rank correlation coefficient. This would have the advantage over the correlation coefficient of being less sensitive to outliers, but would still ignore the time ordering of the observations.

One proposal for a new dissimilarity measure which takes account of the time ordering of the data is to standardise the observations for each day to have a zero mean and a unit variance and to calculate the area between the two standardised curves thus generated.

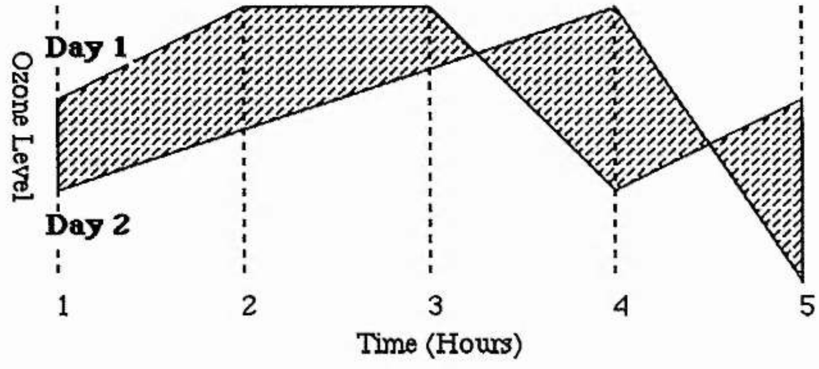


Fig. 5.1

Fig 5.1 illustrates this approach, with the difference between the two sections of curve from day 1 and day 2 represented by the shaded area.

This measure of dissimilarity does indeed look for a difference in shape between the objects, as the curves have been standardised, and the time ordering of the observations is accounted for. For these reasons it has been adopted as a dissimilarity measure for this analysis. A closer investigation of the statistical properties of this dissimilarity measure is made in chapter 6.

Dissimilarity between clusters

There are many possible ways of defining the dissimilarity between two clusters C_i and C_j (d_{ij}) once the measure of dissimilarity between two objects has been chosen. (Gordon 1987)

The focus of this analysis is to obtain estimates of the typical day for each type found. These will be estimated by the cluster means (M_i). The importance of the cluster means to this analysis leads us to define the dissimilarity between two clusters as the dissimilarity between their means. With many standard dissimilarity measures, this would be equivalent to taking the mean of all the pairwise differences between objects in C_i and C_j. With the area dissimilarity measure introduced above, this is not the case and we must calculate the two cluster means at each iteration. This distinction will be discussed in chapter 6.

Choice of clustering algorithm

One possibility for the clustering algorithm is to start with all the objects in the data set divided into some initial classification. At each iteration the algorithm seeks a change in the existing classification consisting of either a move of one object from one cluster to another or a swap of two objects in different clusters such that the greatest possible reduction in some measure of global stress is achieved. This method of seeking the best swap or move continues until no reduction in stress can be achieved.

To implement this algorithm an initial classification of the data set is required. This could be either the end result of some other classification routine, a simple random classification or some subjective allocation of days to clusters, perhaps chosen with reference to data on weather types or some other explanatory variables. This would be undesirable in view of the fact that in this thesis the approach is to consider such data only after analysis.

The fact that the initial classification chosen can affect the end result of a classification by this method (Gordon 1987) is another disadvantage of this approach.
This method can also be computationally intensive for large data sets.

Another approach is to use an agglomerative algorithm This begins with each element considered as a separate cluster of size one. At each iteration the 'closest' two clusters are combined to form a new one and the number of clusters is thus reduced by one. If C_i and C_j have been combined to form C_k, then C_k will be represented by M_k (the mean of cluster k) at the next iteration. This procedure of combining two clusters at each iteration continues until the data set is combined into a single cluster. This algorithm has two primary advantages over the one considered above:

First, there is no need to form an initial classification of the data.

Second, this algorithm is not as computationally intensive as the first. (see below)

For the above reasons, it was decided to implement an agglomerative algorithm based on the area dissimilarity measure introduced above. The statistical properties of this approach are discussed in chapter 6.

Computing time requirements

To investigate the computational requirements of these methods, it is assumed below that there are a total of 366 objects and that there are 4 clusters required.

For the first algorithm it is only possible to evaluate the number of calculations required at a typical iteration, as we do not know how many iterations will be required.

If at some iteration there are N clusters of sizes n_i (i=1...N), then there are a total of

$$(N-1)\sum_{i=1}^{N} n_i \qquad \text{possible moves}$$

and

$$\sum_{i=1}^{N} \sum_{j=1\ j<i}^{N} n_i n_j \qquad \text{possible swaps.}$$

So if the objects have been divided into 4 roughly equal clusters, say of sizes 91,91,92,92, then there are 1098 possible moves and 50233 swaps, leading to a total of 51331 possible changes. Each of these requires the recalculation of 2 cluster means and an average of 183 differences between an object and its cluster mean. This leads to a total of about $1.88\text{x}10^7$ calculations for a single iteration.

However, for the second algorithm we must first calculate a matrix of all pairwise differences. If we have N objects this is of size N^2. As this matrix is symmetric and has all main diagonal elements zero, we need only calculate $\frac{1}{2}$ N(N-1) of these.

At each iteration we combine the two closest clusters into one, calculate one cluster mean, eliminate one row and column of the matrix, and calculate the entries of a new row and column. If we have n clusters after the combination stage, there are n-1 entries to be calculated.

We also know that if we require C clusters at the end of the analysis there are a total of N-C iterations.

In this example, we require an initial matrix consisting of 66795 elements, 362 mean calculations, and

$$\sum_{i=5}^{365} (i-1)=66976 \qquad \text{calculations of matrix elements.}$$

This leads to a total of 1.34×10^5 calculations for the entire classification, considerably fewer than was required for one iteration of the other algorithm.

Implementation

To implement the chosen classification method, a FORTRAN program has been written. (listing in appendix 2)

The program implements the method as follows:

The user is asked to input the number of objects ; the number of observations per object; whether any standardisation is required (mean and or standard deviation); and the range of number of clusters for which output is required. It is also possible to enter a weight function to be used when calculating the differences, to give more weight to certain parts of the day should that be felt appropriate.

For each day in turn, the vectors of interpolated values $\underline{y}_i$ are entered, and any standardisation is carried out. The standardisation used is (Observation value - mean for that object) / (standard deviation of that object), and is carried out before any clustering occurs.

To begin with there are N objects. The program first calculates the lower triangular matrix of dissimilarities, using the area dissimilarity measure and the weight function if entered. With n observations in each object, there are a set of n-1 weights and the area between the i^{th} and $(i+1)^{st}$ observations of 2 objects is multiplied by the i^{th} weight before being summed.

Once the data entry is complete, the program implements the agglomerative algorithm discussed above. At each iteration if clusters C_i and C_j are to be joined the program will output i,j and d_{ij} and in addition, for the range of classifications specified, a list of which objects have been allocated to each cluster along with the cluster means M_i.

5.4: Results

Each object for classification, whether it represents a whole day or half a day, has been standardised to have a zero mean and a unit variance before analysis.

Clustering Results

Table 5.1 shows a summary of the output generated by the classification program for the whole day data set. Although output has been given for g=10 to g=2 (where g=k refers to the k cluster classification), at each stage from g=10 to g=3 there are two large clusters and a number of small clusters, some containing only one day, implying the existence of two main types of whole day curve.

Table 5.1: 1998 Bush Whole Day Classification

Number of Clusters	Sizes
10	1,1,1,1,1,1,9,1,82,238
9	1,1,2,1,1,9,1,82,238
8	1,1,2,1,9,1,83,238
7	1,1,2,1,1,92,238
6	1,1,2,1,93,238
5	1,1,2,94,238
4	1,2,94,239
3	1,96,239
2	1,335

To have as many days as possible allocated to one of these two types the classification chosen is that from g=3, taking clusters two and three as the two types. Figs 5.2 and 5.3 show the means for these two clusters. There are a total of 96 days allocated to the first cluster and 239 days allocated to the second.

Fig 5.2: Whole day - Mean of cluster 2

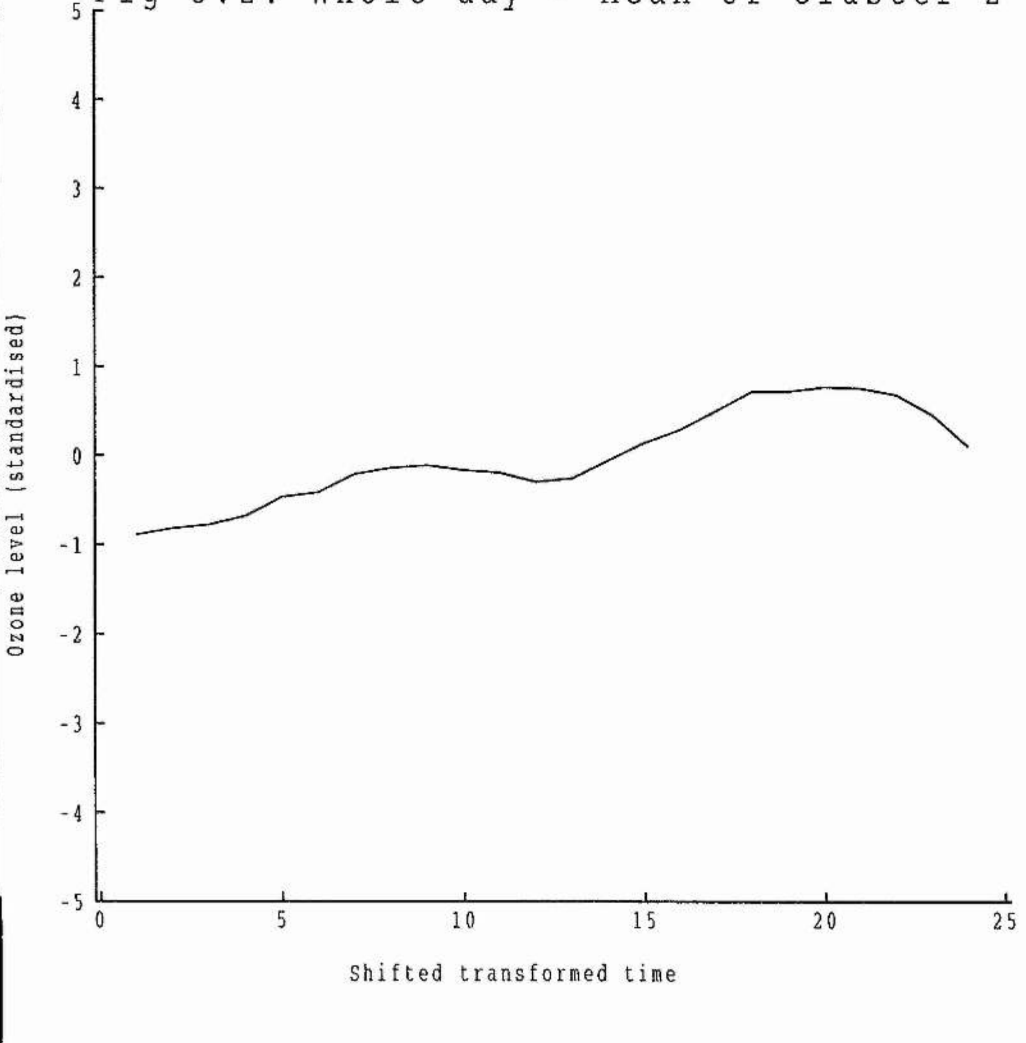


Fig 5.3: Whole day - Mean of cluster 3

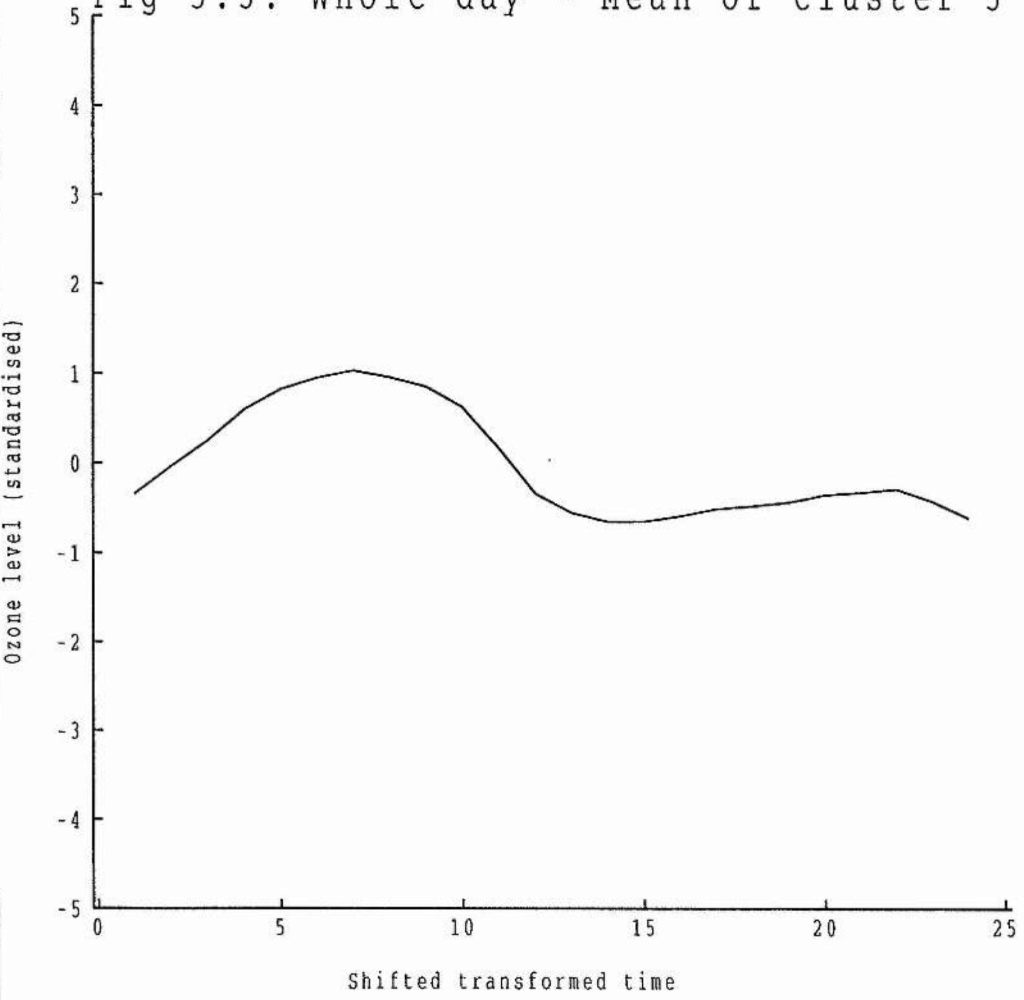


Table 5.2 shows a similar summary for the daytime only data set. A decision on the number of clusters to be taken is not so clear cut as for table 5.1. The choice is between g=10 with four main clusters, g=4 with three or g=2 with two. The first choice has been taken, the four groups being of sizes 74,172,36 and 45 respectively, with a total of 9 days unallocated. Figs 5.4,5.5,5.6 and 5.7 show their means.

Table 5.2: 1988 Bush Daytime Classification

Number of Clusters	Sizes
10	1,1,2,1,3,1,74,172,36,45
9	1,1,2,1,3,1,172,110,45
8	1,2,1,3,1,172,110,46
7	1,2,1,3,172,110,47
6	1,2,1,172,113,47
5	3,1,172,113,47
4	3,172,114,47
3	3,114,219
2	117,219

Table 5.3 shows a comparable summary for the night time only data set. The choice of classification is between g=6 with five main clusters, g=5 with four or g=3 with three. The first choice has been taken, the five groups being of sizes 14,81,95,124 and 21 respectively. Figs 5.8,5.9,5.10,5.11 and 5.12 show their means.

Fig 5.4: Daytime - Mean of cluster 7

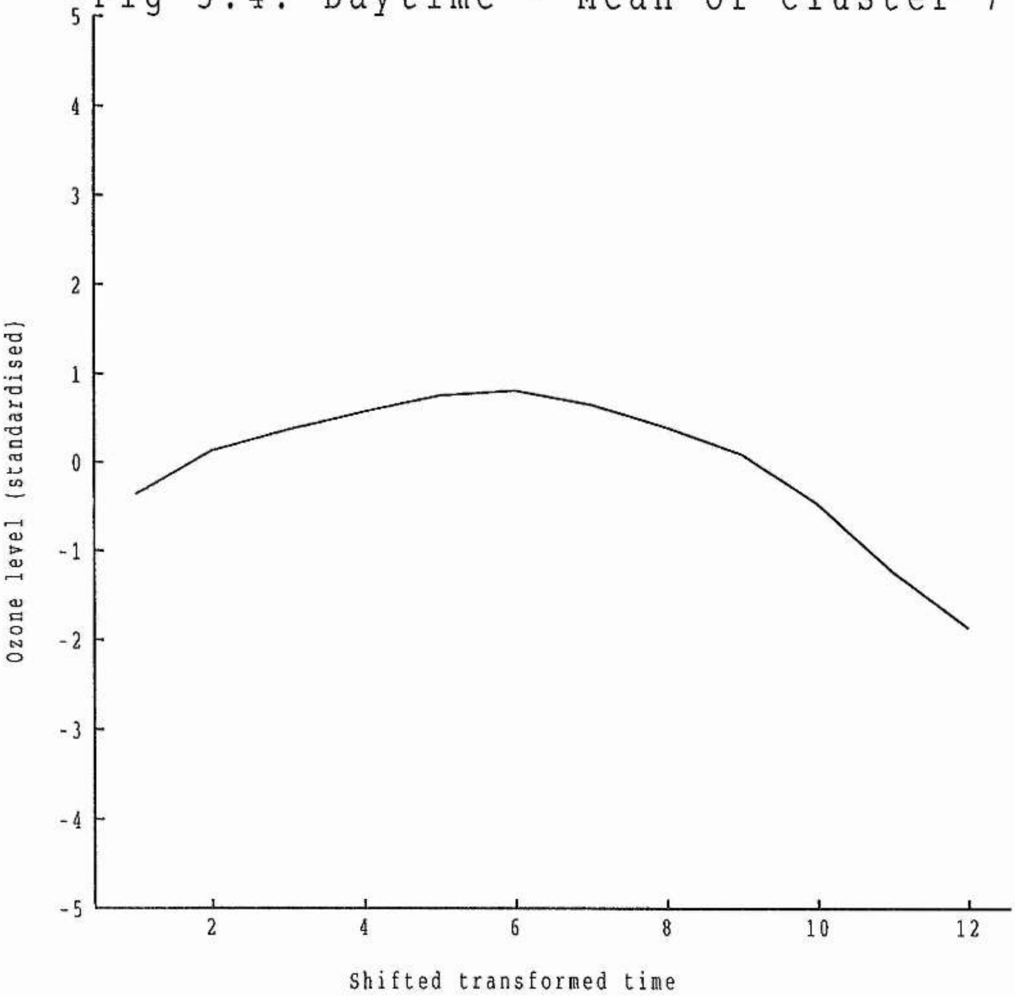


Fig 5.5: Daytime - Mean of cluster 8

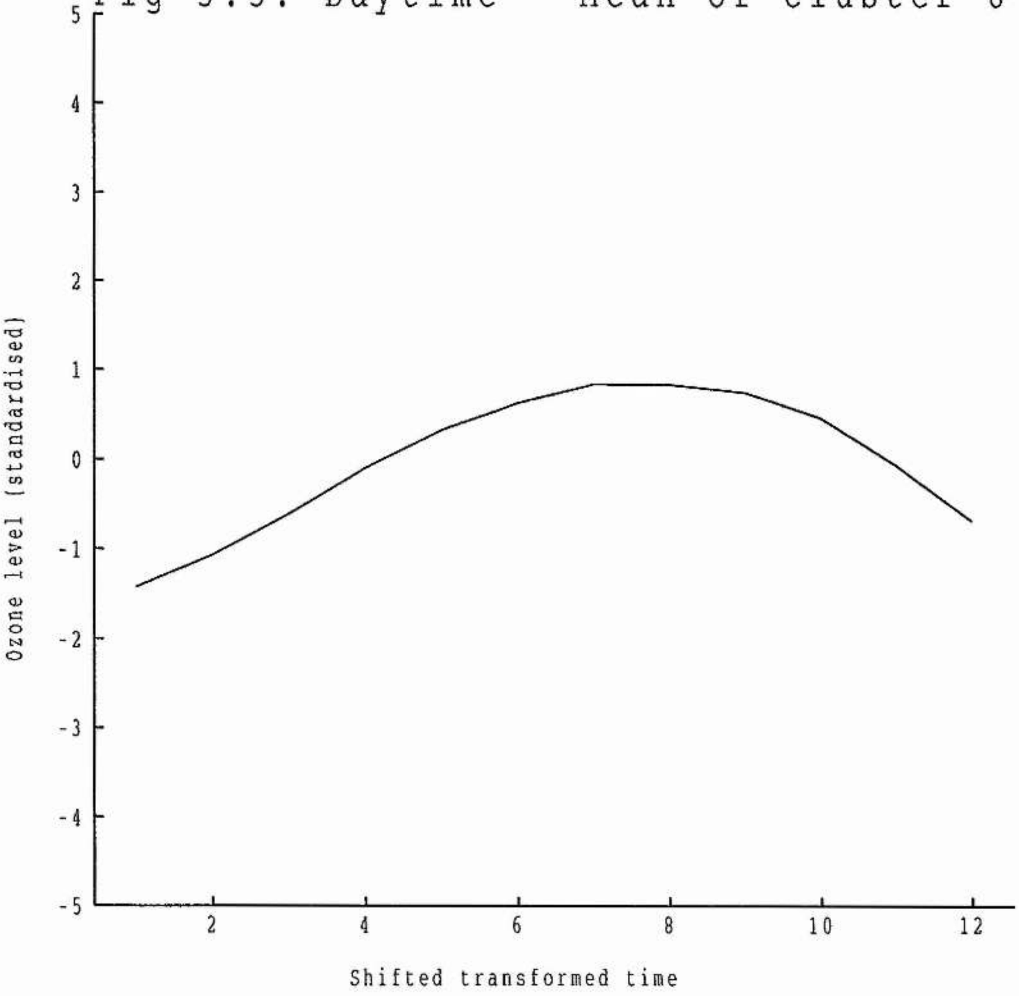


Fig 5.6: Daytime - Mean of cluster 9

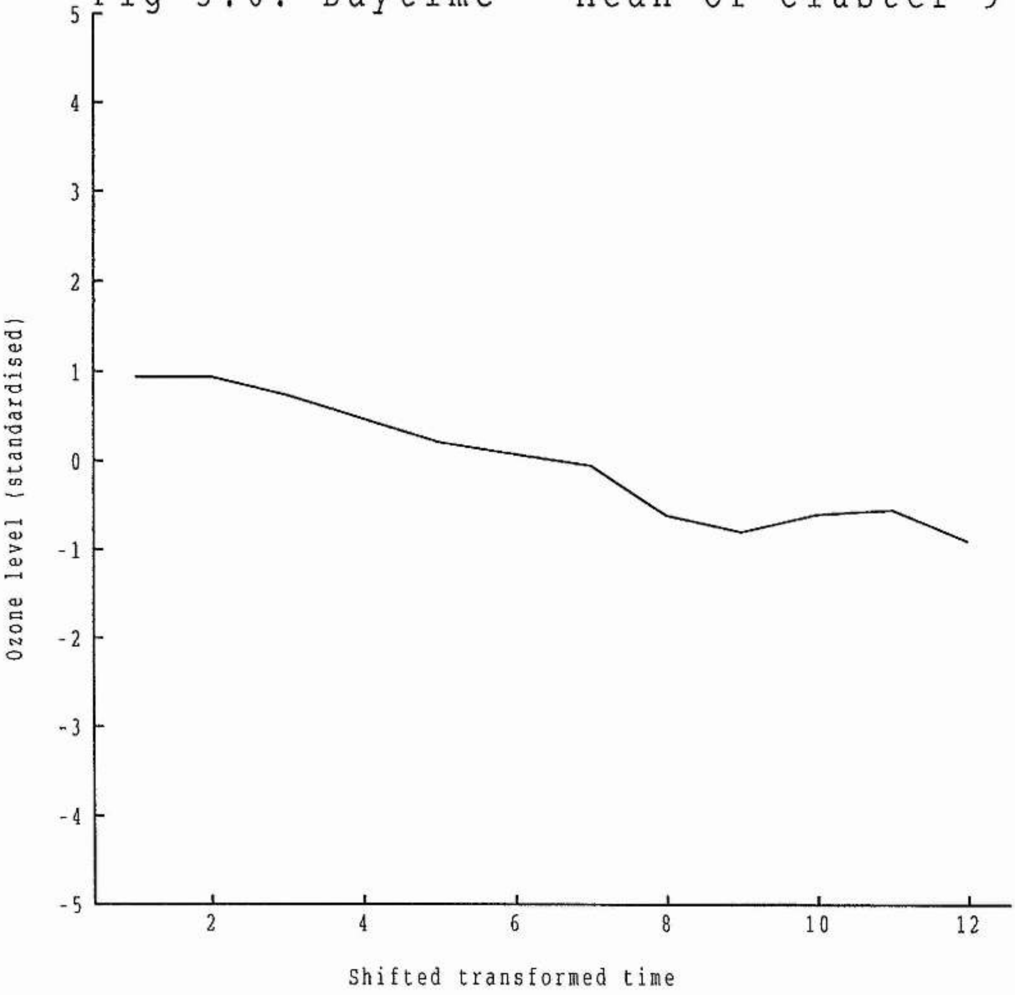


Fig 5.7: Daytime - Mean of cluster 10

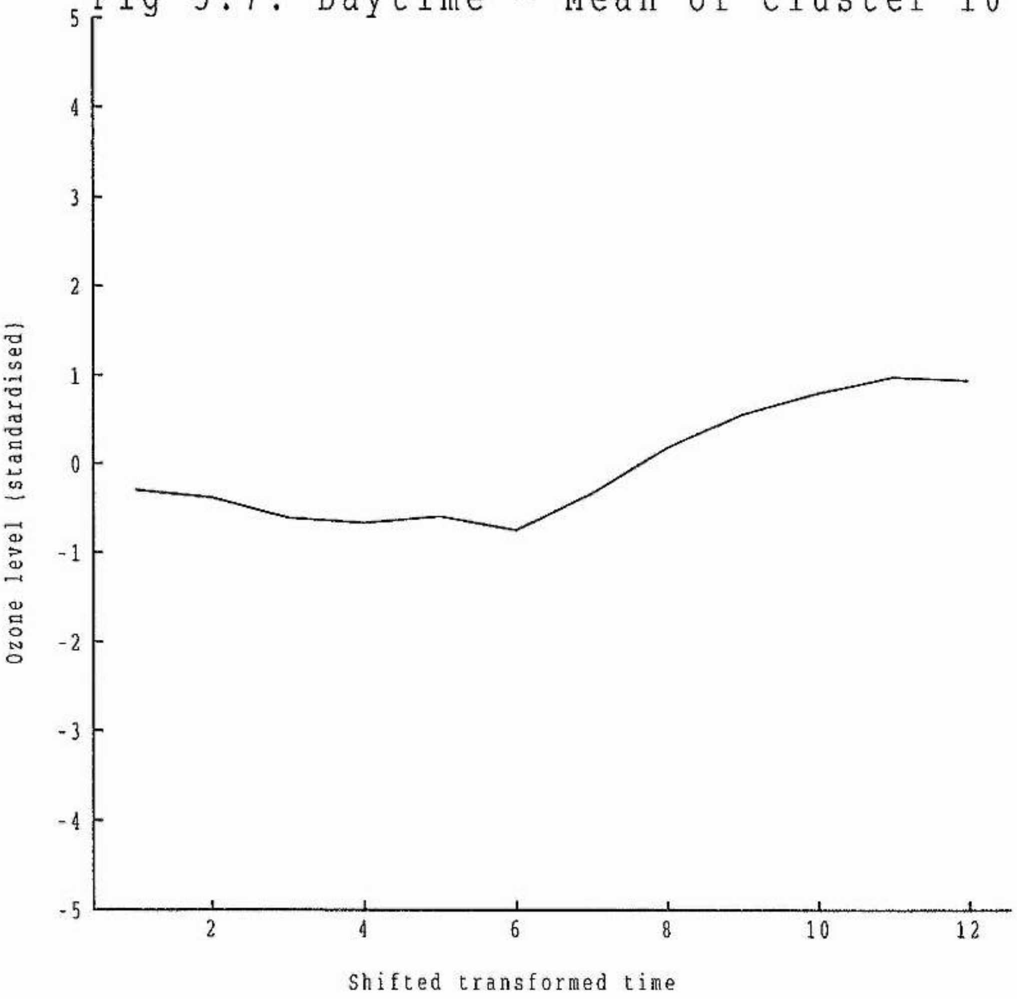


Fig 5.8: Night time - Mean of cluster 2

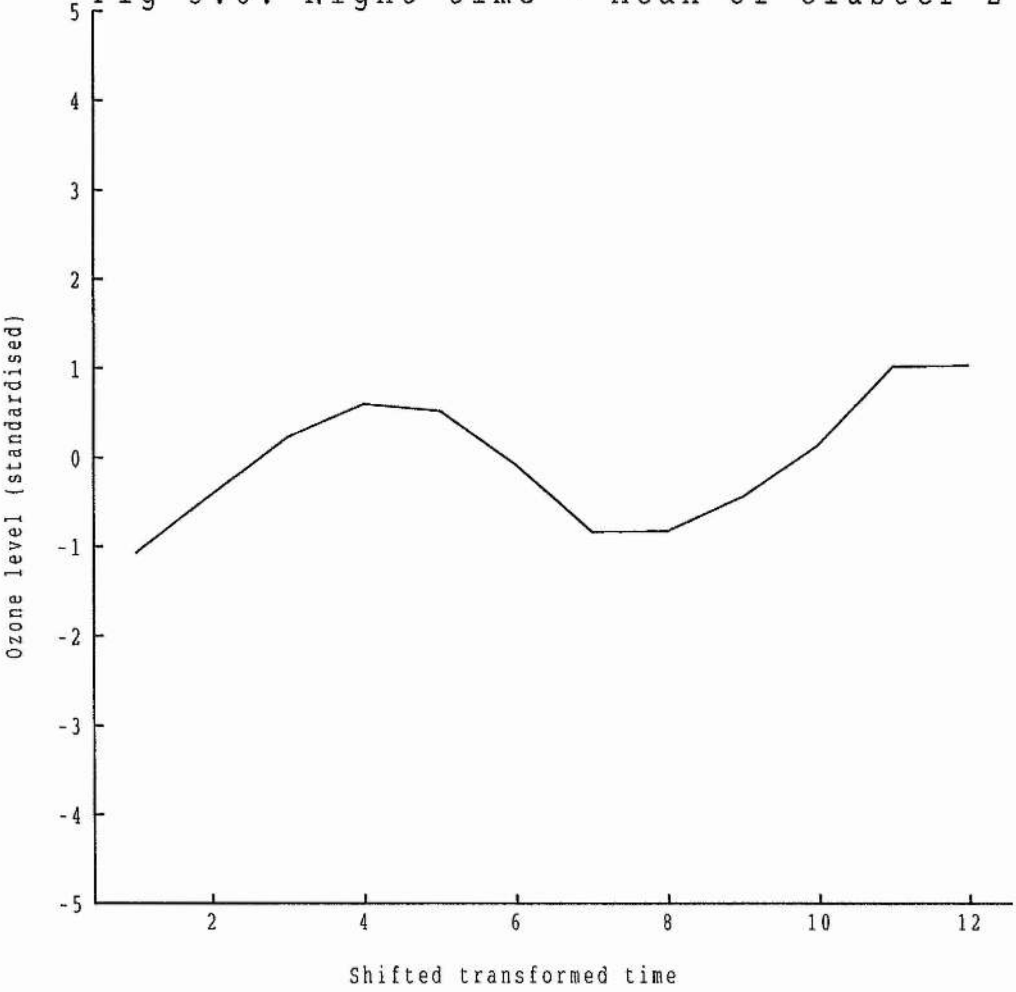


Fig 5.9: Night time - Mean of cluster 3

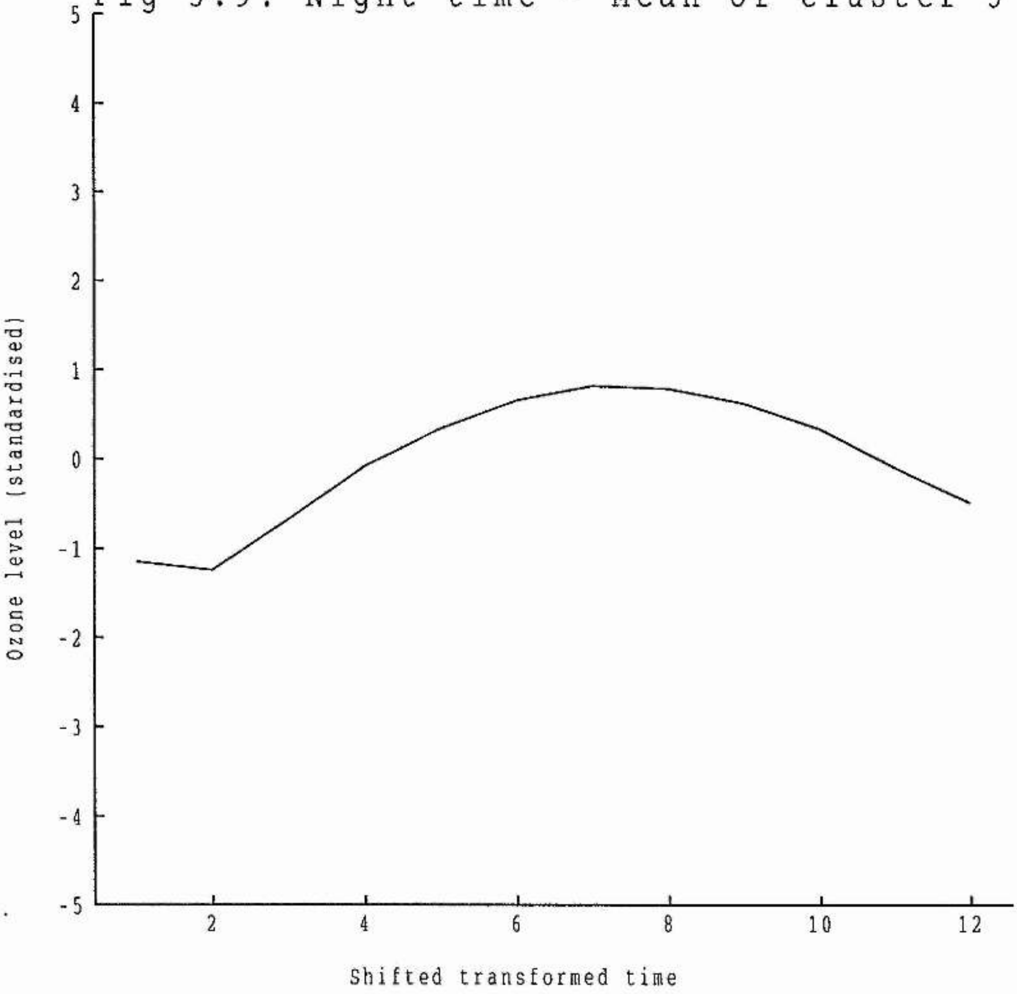


Fig 5.10: Night time - Mean of cluster 4

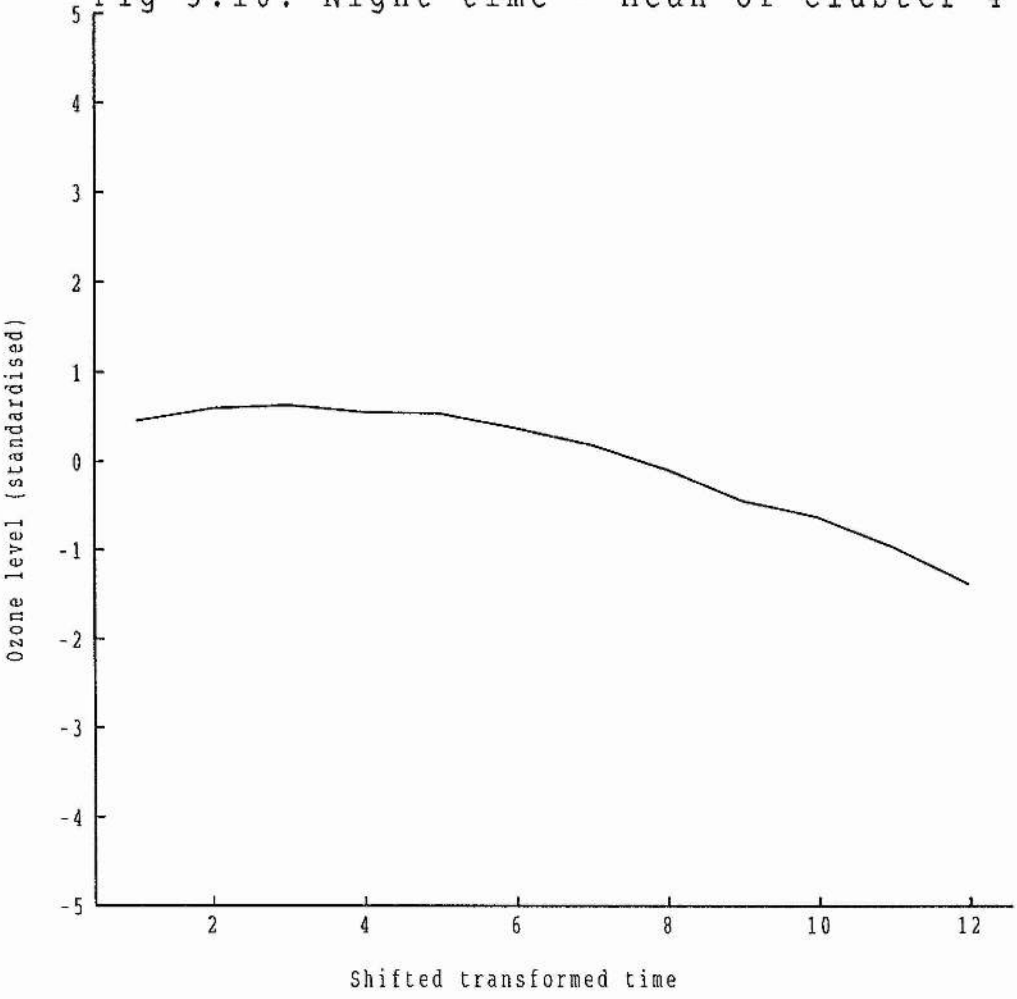


Fig 5.11: Night time - Mean of cluster 5

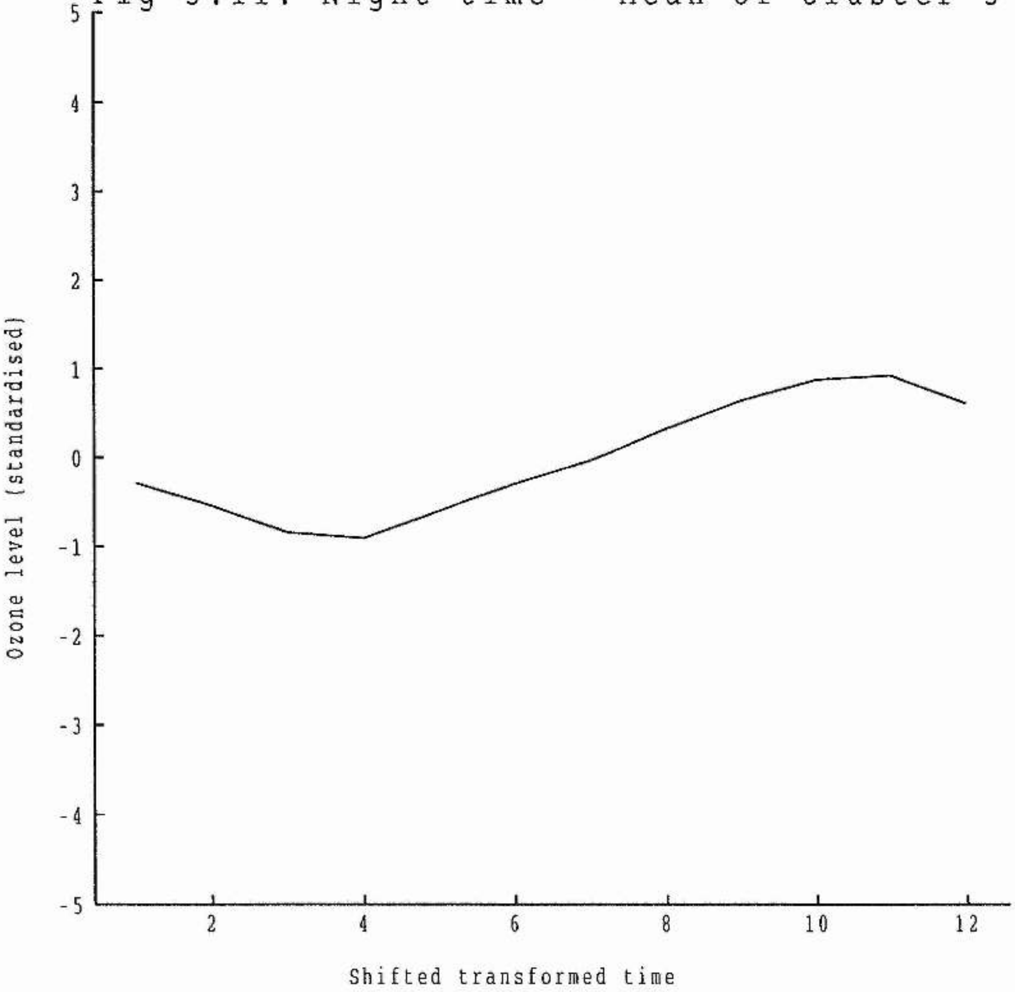


Fig 5.12: Night time - Mean of cluster 6

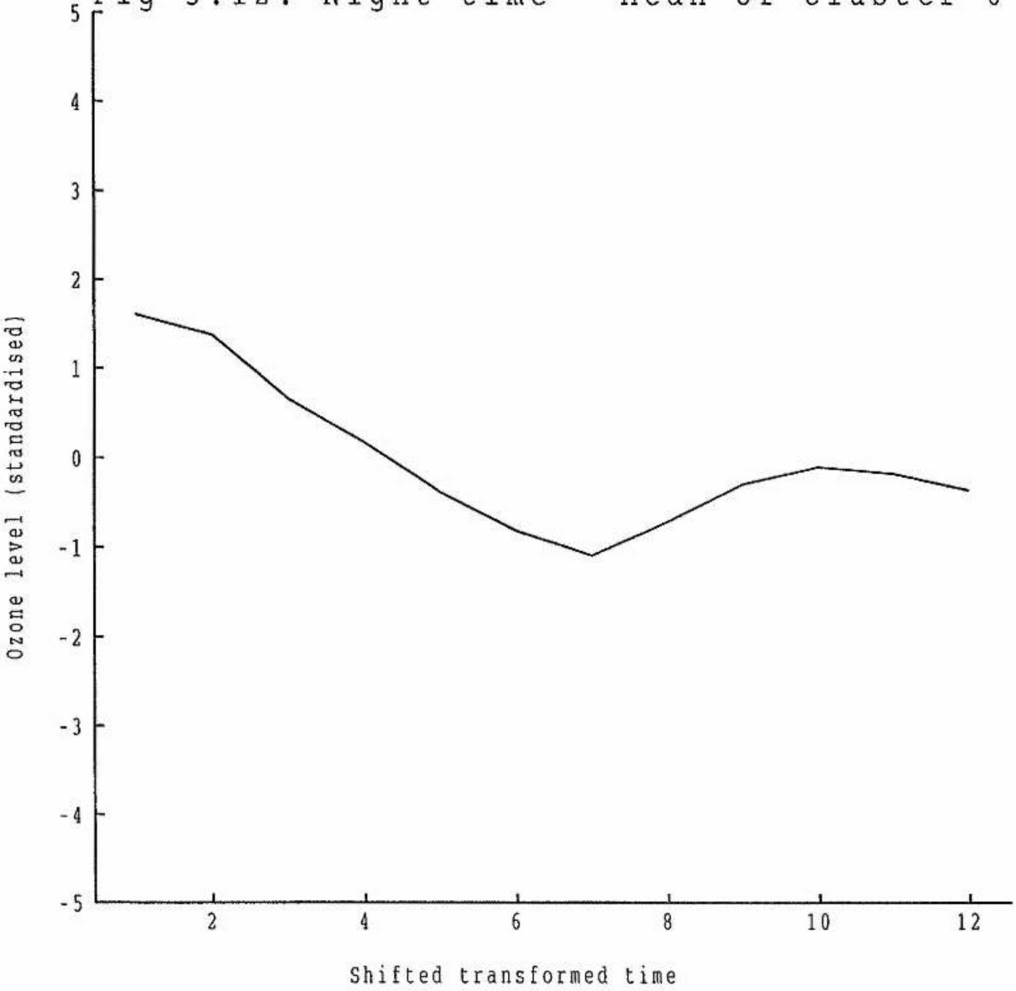


Table 5.3: 1988 Bush Night Time Classification

Number of Clusters	Sizes
10	1,1,1,9,1,13,81,85,123,21
9	1,1,1,1,13,81,94,123,21
8	1,1,1,13,81,95,123,21
7	1,1,14,81,95,123,21
6	1,14,81,95,124,21
5	1,14,95,205,21
4	1,14,205,116
3	14,206,116
2	206,130

Figs 5.13,5.14 and 5.15 show principal coordinate plots of the whole day, daytime and night time data sets respectively. Each object is represented by a number indicating which cluster that object has been allocated to. None of these data sets are visibly divided into disjoint clusters, but the clusters found by the above method represent different regions of the data sets. If the assumption that the type of ozone curve is determined by the weather for each day is correct, this lack of disjoint clusters would be anticipated as the type of weather observed each day comes from a continuum of possible patterns.

This classification method has discovered two main types of whole day curve, four types of daytime curve and five types of night time curve. The types of curve will be referred to as W_1,W_2,D_1-D_4 and N_1-N_5 for the whole day, daytime and night time curves respectively.

Fig 5.13: Whole day clustering result g=3

1: Cluster 1, size 1
2: Cluster 2, size 96
3: Cluster 3, size 239

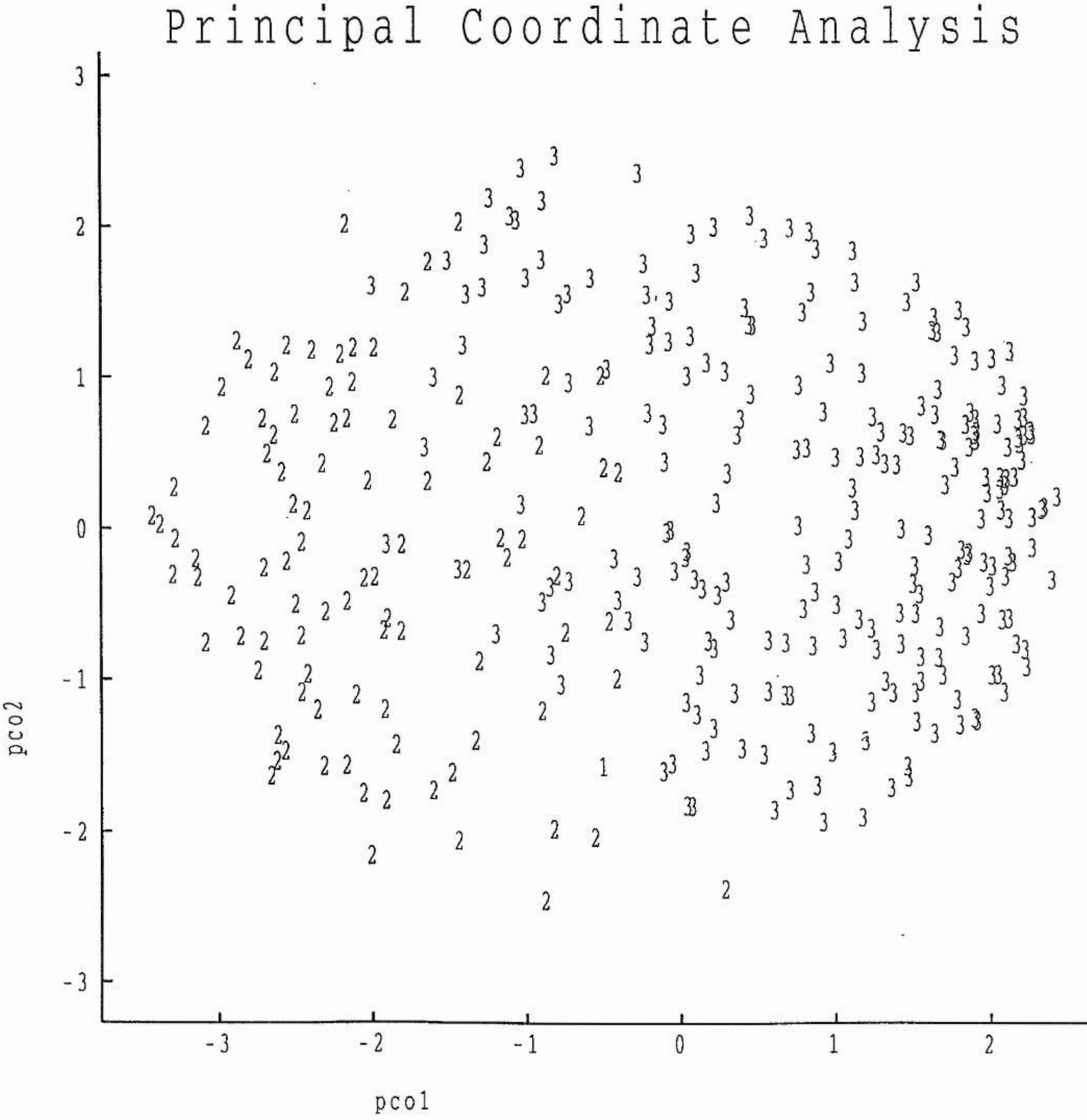


Fig 5.14: Daytime clustering result g=10

1: Clusters 1-6, sizes 1,1,2,1,3,1
2: Cluster 7, size 74
3: Cluster 8, size 172
4: Cluster 9, size 36
5: Cluster 10, size 45

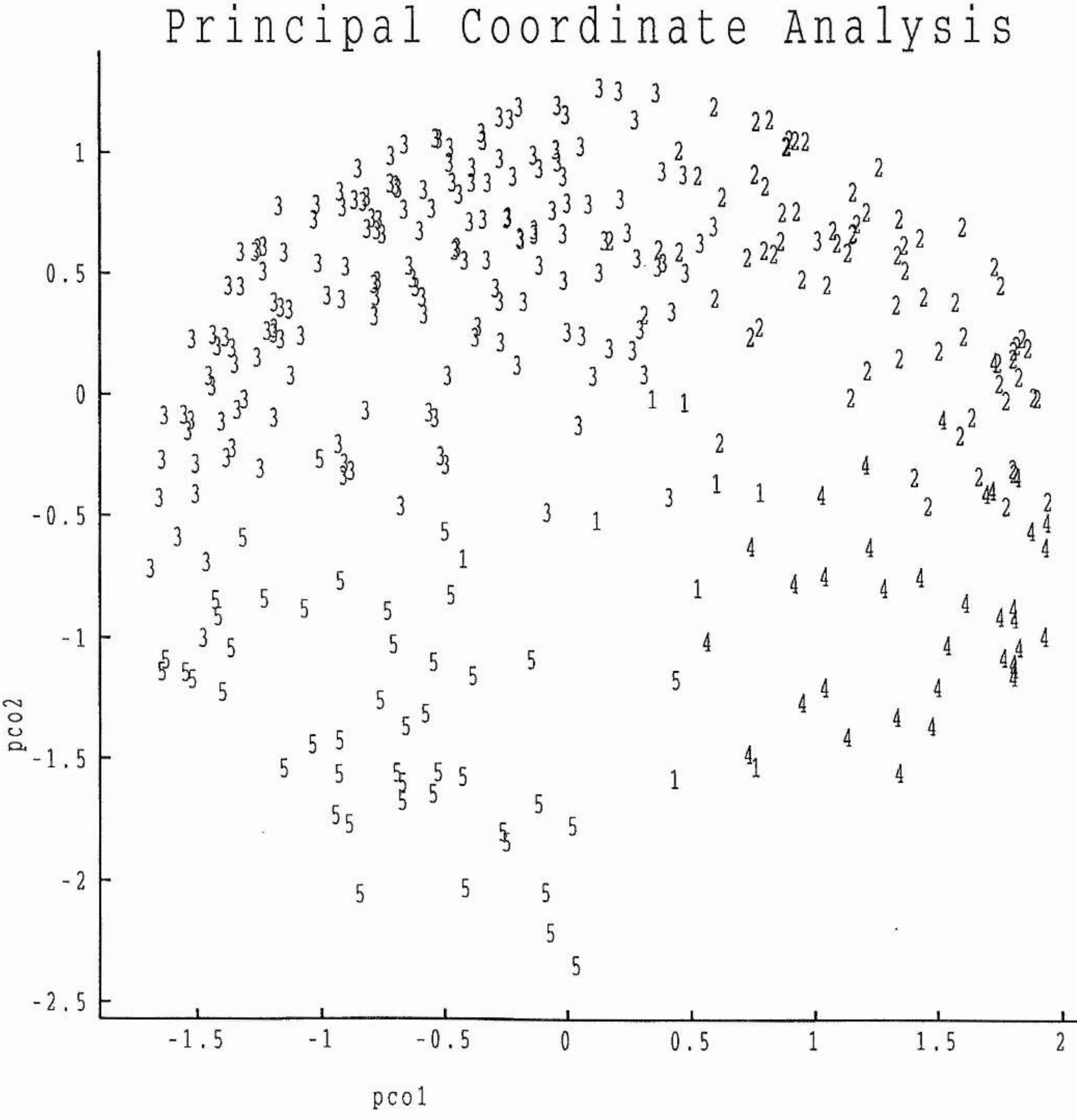
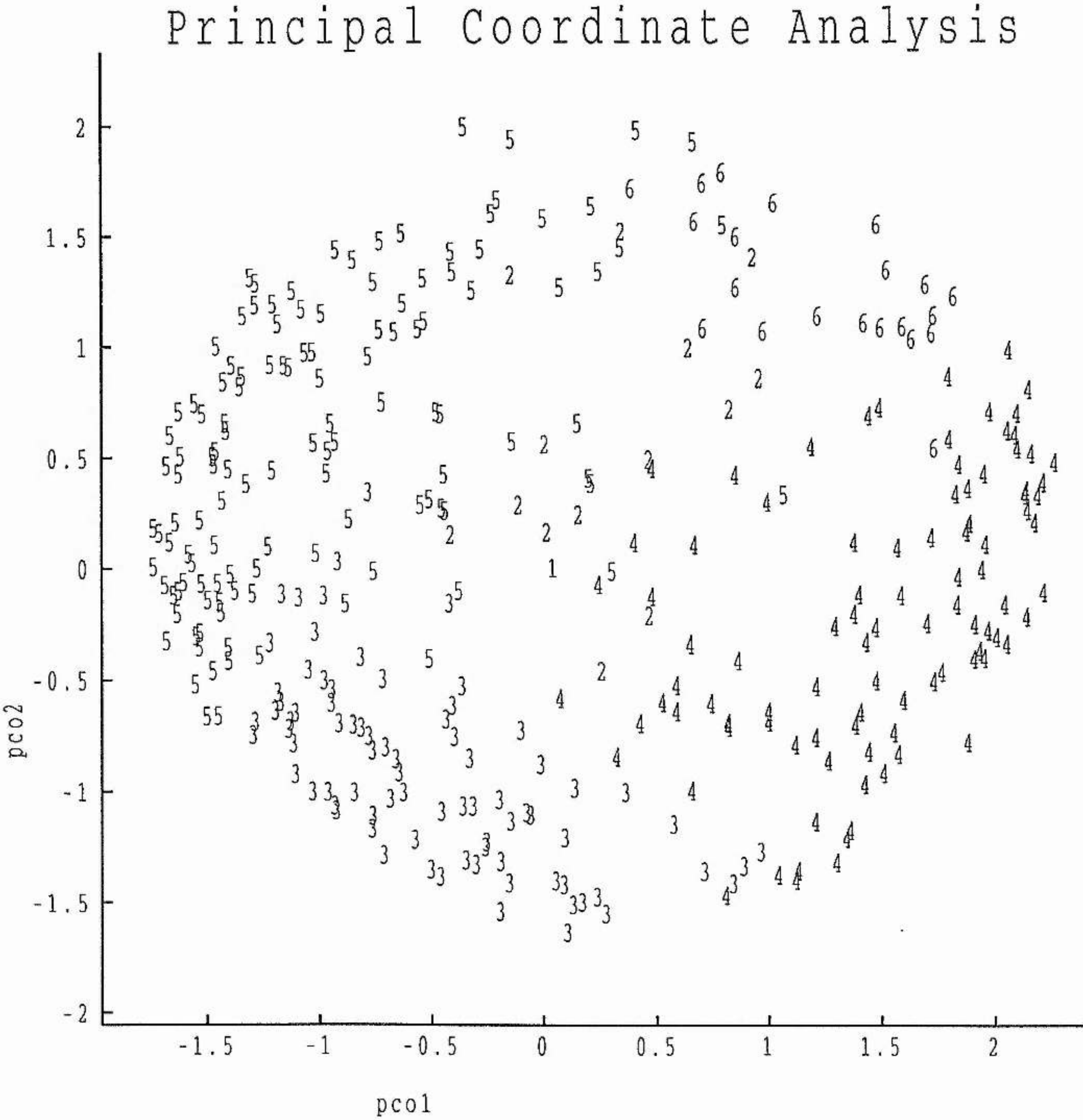


Fig 5.15: Night time clustering result g=6

1: Cluster 1, size 1
2: Cluster 2, size 14
3: Cluster 3, size 81
4: Cluster 4, size 95
5: Cluster 5, size 124
6: Cluster 6, size 21



Mean values within clusters

The mean ozone level for each object has been extracted and an analysis of these with respect to the cluster each object has been allocated to has been carried out.

Table 5.4 shows summary statistics of the distributions of the mean levels of the days placed in the two main clusters for the whole day data set. It is easily seen that there is no significant difference in the mean levels of days placed in either of these two clusters.

Table 5.4: Distribution of Mean Ozone Levels for W1 and W2

	Number of Days	Mean	Median	StDev	Min	Max	Q1	Q3
W1	96	26.282	26.950	7.232	7.240	53.250	21.460	31.237
W2	239	27.130	27.950	7.353	4.730	54.120	22.460	31.340

Table 5.5 shows summary statistics of the distributions of the mean levels of the daytime sections placed in the four main clusters for the daytime only data set. Also shown are the results of a Kruskal-Wallis nonparametric test, and the relevant pairwise Mann-Whitney tests. The mean levels for the different types of daytime curve are significantly different. The means for D_2 are significantly higher than the others, the means for D_4 are significantly lower than the others and the means for D_1 and D_3 are not significantly different from each other.

Table 5.5: Distribution of Mean Ozone Levels for D1 - D4

	Number of Days	Mean	Median	StDev	Min	Max	Q1	Q3
D1	74	28.189	27.985	7.870	6.600	50.780	23.625	32.385
D2	172	30.504	30.700	8.015	4.840	58.300	26.548	35.177
D3	36	28.580	30.700	7.220	12.120	41.260	25.12	33.060
D4	45	23.380	25.480	7.290	6.540	34.310	17.050	29.670

Kruskal Wallis Test

Level	N	Median	Av. Rank	Z Value
1	74	27.99	151.3	-1.31
2	172	30.70	184.4	4.11
3	36	30.70	170.0	0.40
4	45	25.48	102.1	-4.73
Overall	327		164.0	

H=28.80, d.f.=3, p=0.000 Highly significant difference

Table 5.6 shows summary statistics of the distributions of the mean levels of the night time sections placed in the five main clusters for the night time only data set. As above, the results of the Kruskal-Wallis and Mann-Whitney tests are also shown. The mean levels for N_2 are slightly higher than the other four, but there is less of a difference observed than was for the daytime sections.

Table 5.6: Distribution of Mean Ozone Levels for N1 - N5

	Number of Days	Mean	Median	StDev	Min	Max	Q1	Q3
N 1	14	27.670	26.490	5.650	19.010	38.590	23.850	30.120
N 2	81	26.903	28.440	8.181	6.050	54.610	22.160	32.035
N 3	95	24.783	25.400	9.078	3.200	49.940	18.880	30.390
N 4	124	24.286	25.750	8.766	2.950	48.200	18.385	30.177
N 5	21	22.190	25.280	7.350	7.390	32.880	13.880	27.880

Kruskal Wallis Test

Level	N	Median	Av. Rank	Z Value
1	14	26.49	190.9	0.90
2	81	28.44	189.5	2.30
3	95	25.40	163.7	-0.51
4	124	25.75	160.5	-1.09
5	21	25.28	133.5	-1.68
Overall	335		168.0	

H=8.38, d.f.=4, p=0.079 Test significant at 7.9%

Variances within clusters

The variances of the observations making up each object have also been extracted and a similar analysis to that of the means has been carried out.

Table 5.7 shows summary statistics of the distribution of the variances of the days placed in the two main clusters for the whole day data set. There is a significant difference between the variances of days placed in either of these two clusters. The variances of W_2 are higher than those of W_1.

Table 5.7: Distribution of Variances for W1 and W2

	Number of Days	Mean	Median	StDev	Min	Max	Q1	Q3
W1	96	33.97	23.67	36.47	1.01	219.19	8.86	45.80
W2	239	53.41	33.82	53.35	1.82	269.12	14.16	76.14

Mann-Whitney Test

95% Confidence Interval for difference between means is (-16.34,-3.62)

Test of Mean1=Mean2 vs Mean1 n.e. Mean2 is significant at 0.12%

Table 5.8 shows summary statistics of the distributions of the variances of the daytime sections placed in the four main clusters for the daytime only data set. There are significant differences between the variances of daytime sections placed in these four clusters.

Table 5.8: Distribution of Variances for D1 - D4

	Number of Days	Mean	Median	StDev	Min	Max	Q1	Q3
D1	74	47.26	32.53	55.80	0.35	340.14	10.65	64.74
D2	172	38.59	20.23	48.69	0.52	326.58	6.89	51.55
D3	36	17.07	6.78	20.70	0.29	82.96	2.62	27.54
D4	45	20.92	11.73	24.64	1.85	113.57	4.02	27.86

Kruskal Wallis Test

Level	N	Median	Av. Rank	Z Value
1	74	32.526	195.8	2.74
2	172	20.228	178.9	2.01
3	36	6.777	120.0	-3.17
4	45	11.732	139.4	-2.16
Overall	327		164.0	

H=27.59, d.f.=3, p=0.000 Highly significant difference

Table 5.9 shows summary statistics of the distributions of the variances of the night time sections placed in the five main clusters for the night time only data set. There is no significant difference between the variances of the night time sections placed in any of these clusters.

Table 5.9: Distribution of Variances for N1 - N5

	Number of Days	Mean	Median	StDev	Min	Max	Q1	Q3
N 1	14	17.86	15.02	13.95	1.46	51.62	8.52	23.82
N 2	81	24.79	17.62	25.41	0.26	107.10	6.07	32.18
N 3	95	24.54	17.46	25.02	0.26	116.94	6.42	35.09
N 4	124	21.03	11.41	24.42	0.46	117.29	5.27	23.72
N 5	21	17.11	11.00	14.04	0.96	49.34	7.28	32.34

Kruskal Wallis Test

Level	N	Median	Av. Rank	Z Value
1	14	15.02	169.1	0.02
2	81	17.62	178.8	1.09
3	95	17.46	176.2	0.91
4	124	11.41	157.4	-1.61
5	21	11.00	160.0	-0.41
Overall	335		168.0	

H=3.30, d.f.=4, p=0.654 Test not significant

Difference in mean level

Although no difference has been discovered between the mean levels for whole days in either of the two clusters, it was hypothesised that there might be a correlation between the change in mean level from day to day and the types of these two days. To investigate this the change in mean level between each pair of consecutive days of the year was calculated. These differences have been split into four groups corresponding to the four possible combinations of types of first day and second day. A summary of these differences is shown in table 5.10. Results from a Kruskal-Wallis test are also shown.

There is no statistically significant difference between these four sets of differences.

Table 5.10: Difference in mean level from one day to the next

Shape					
Day 1	Day 2	Number	Mean	Median	StDev
1	1	37	1.70	1.75	6.47
1	2	59	-0.676	-1.364	6.109
2	1	58	1.107	1.065	6.922
2	2	179	-0.467	-0.374	7.515

Kruskal Wallis Test

Level	N	Median	Av. Rank	Z Value
1	37	1.75	191.5	1.56
2	59	-1.36	155.9	-1.05
3	58	1.06	185.5	1.51
4	179	-0.37	162.0	-1.21
Overall	333		167.0	

H=6.22, d.f.=3, p=0.184 Test not significant

Independence of day/night shapes

Table 5.11 summarises the results of an investigation of the independence of the type of day section and night section for the same day. The test is a χ^2 contingency table type test . The type of night time curve is independent of the type of daytime curve for the day and night sections of the same day. This result implies that a whole day classification may be inappropriate, as if the shape of the daytime curve is independent of the shape of the night time curve for the same day the two should be treated separately.

Table 5.11: Analysis of Shape Ocurrences Whithin Same Whole Day

Day Shape	Night Shape	Number	Expected Number	$(O-E)^2/E$
1	1	5	3.094	1.174
1	2	17	17.830	0.039
1	3	18	20.924	0.409
1	4	29	27.260	0.111
1	5	5	4.642	0.028
2	1	6	7.202	0.201
2	2	44	41.497	0.151
2	3	52	48.699	0.224
2	4	61	63.446	0.096
2	5	8	10.803	0.727
3	1	1	1.506	0.170
3	2	9	8.678	0.012
3	3	10	10.184	0.003
3	4	14	13.268	0.040
3	5	2	2.259	0.032
4	1	2	1.889	0.006
4	2	8	10.887	0.766
4	3	12	12.777	0.047
4	4	17	16.646	0.008
4	5	6	2.834	3.537
			Total	7.781

Test 7.781 as $\chi^2_{(19)}$ gives p=0.988

i.e. Absolutely no evidence that day/night shapes not independent.

5.5: Discussion

Whole day shapes

W_1 represents a day where there is little variation in the observed ozone level during the daytime relative to the variation at night. One possible interpretation of this type of day is a day where there has been little photochemical production of ozone and the variation observed is due to a change in the rate of ozone depletion. The weather for such a day would normally be windy and overcast. If the weather during a day is overcast, there would be little sunlight at ground level and hence the photochemical production of ozone would be minimal. Also, if there is a reasonable amount of wind during the day, the air at the recording site would be constantly being replaced with new air. If there is a constant concentration of ozone present in this new air, and we assume that this replenishment occurs at some fairly constant rate, the observed cycle can be explained as a change in the rate of removal of ozone from this new air. Ozone is removed from the air by NO, the level of which varies during the day, as man made emissions vary with the time of day and particularly they decrease at night. If there is sufficient NO present during the daytime to deplete the ozone from the air, and to keep this concentration at a low level, the observed ozone level at a stationary site will be fairly constant. During the night time, as there will be a reduction in the levels of NO, there will be a rise in ozone levels as this depletion rate decreases. This type of day would be more prevalent during the winter, and also at sites which are more heavily polluted with primary pollutants, such as urban or industrial sites.

W_2 represents a day where there is little variation in the observed ozone level during the night time relative to the variation during the daytime. This can be interpreted as a day where there has been significant photochemical production of ozone, leading to the observed variation of ozone during the daytime which dominates the observed pattern. The weather for such a day would typically be clear and sunny, and as such would be more prevalent during the summer. There is still an observable cycle in the level of ozone during the night, which again can be explained by a change in the rate of ozone depletion. Such a day would also be more prevalent at sites which are less heavily polluted with primary pollutants, such as rural sites.

Weighted average of whole day shapes

This interpretation of the whole day shapes suggests that W_1 should be more prevalent during the winter and W_2 more prevalent during the summer. Table 5.12 shows a breakdown of the classification for the whole day data by month for 1988 and confirms this.

Table 5.12: Frequencies of W1 and W2 for each Month

Month	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
W1	11	15	4	6	4	2	8	1	6	16	11	12
W2	19	12	24	24	21	27	20	29	20	15	14	14

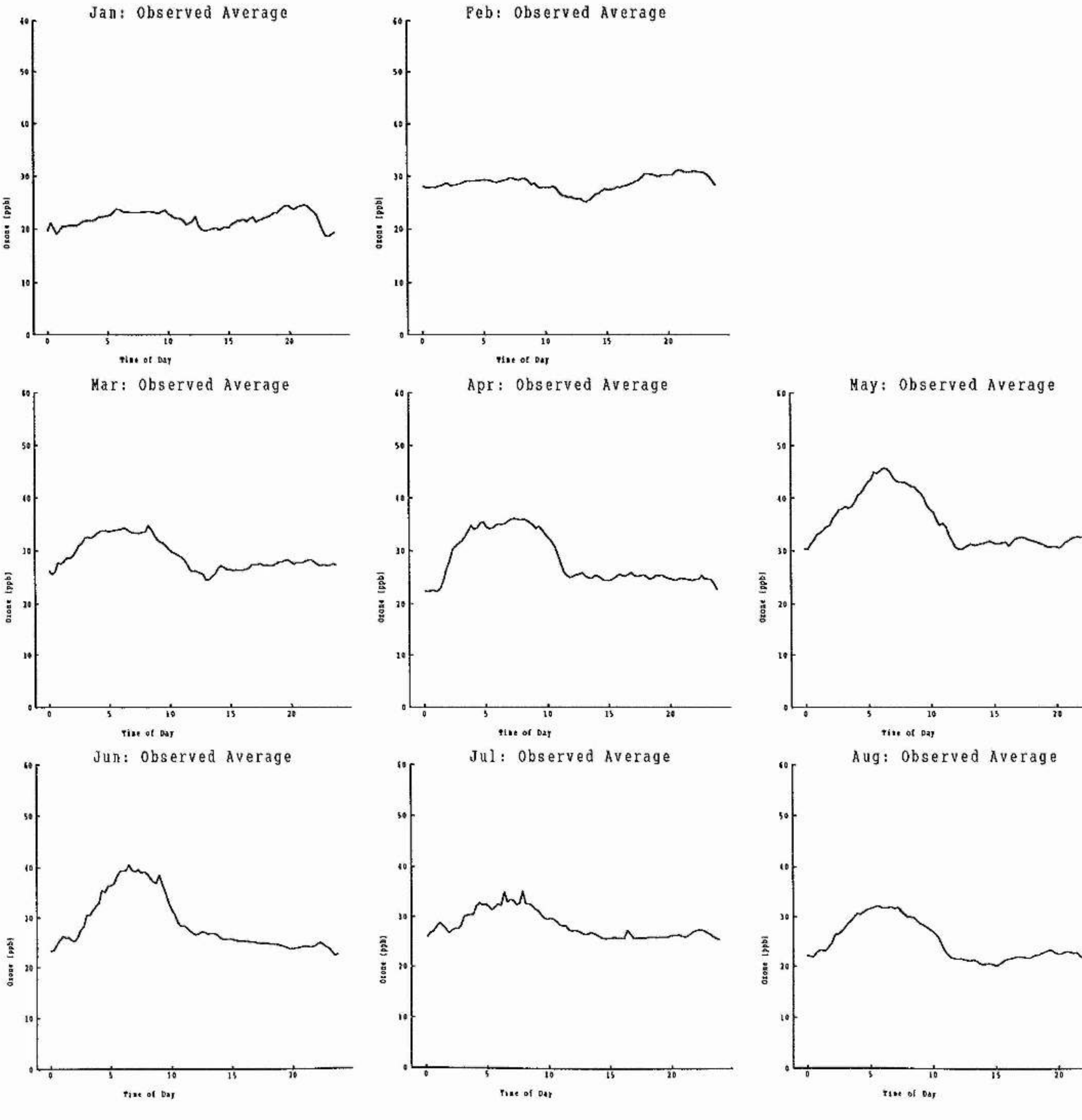
In the previous chapter, it has been shown that the average diurnal cycle varies from month to month. The main observable difference between the averages for each month was a change in the levels of daytime variation relative to the levels of night time variation. To compare results from these monthly averages with the results from this chapter, a weighted average consisting of the two whole day shapes in the proportions indicated in table 5.12 has been calculated for each month.

Fig 5.16 shows these weighted averages next to the average ozone curve for each month calculated in the previous chapter. The weighted averages calculated follow a very similar pattern to the observed monthly averages. Thus the proportion of W_1 to W_2 in each month is sufficient to explain much of the variation observed in the monthly averages.

Daytime shapes

D_1 and D_2 are superficially similar, both having a single 'hump'. The difference between them is the time of the maximum level, D_1 having its maximum about two hours before D_2.

Fig 5.16 (part 1): Weighted average of whole day shapes


Jan: Observed Average

Feb: Observed Average

Mar: Observed Average

Apr: Observed Average

May: Observed Average

Jun: Observed Average

Jul: Observed Average

Aug: Observed Average

Fig 5.16 (part 2): Weighted average of whole day shapes

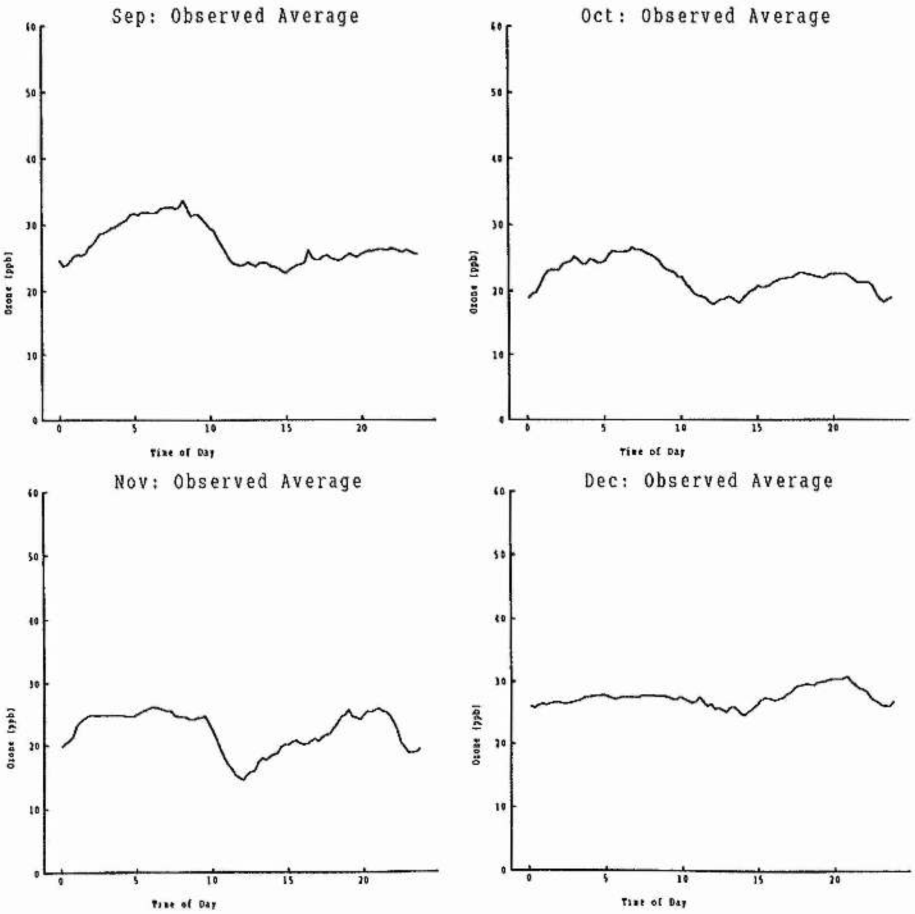


Fig 5.16 (part 3): Weighted average of whole day shapes

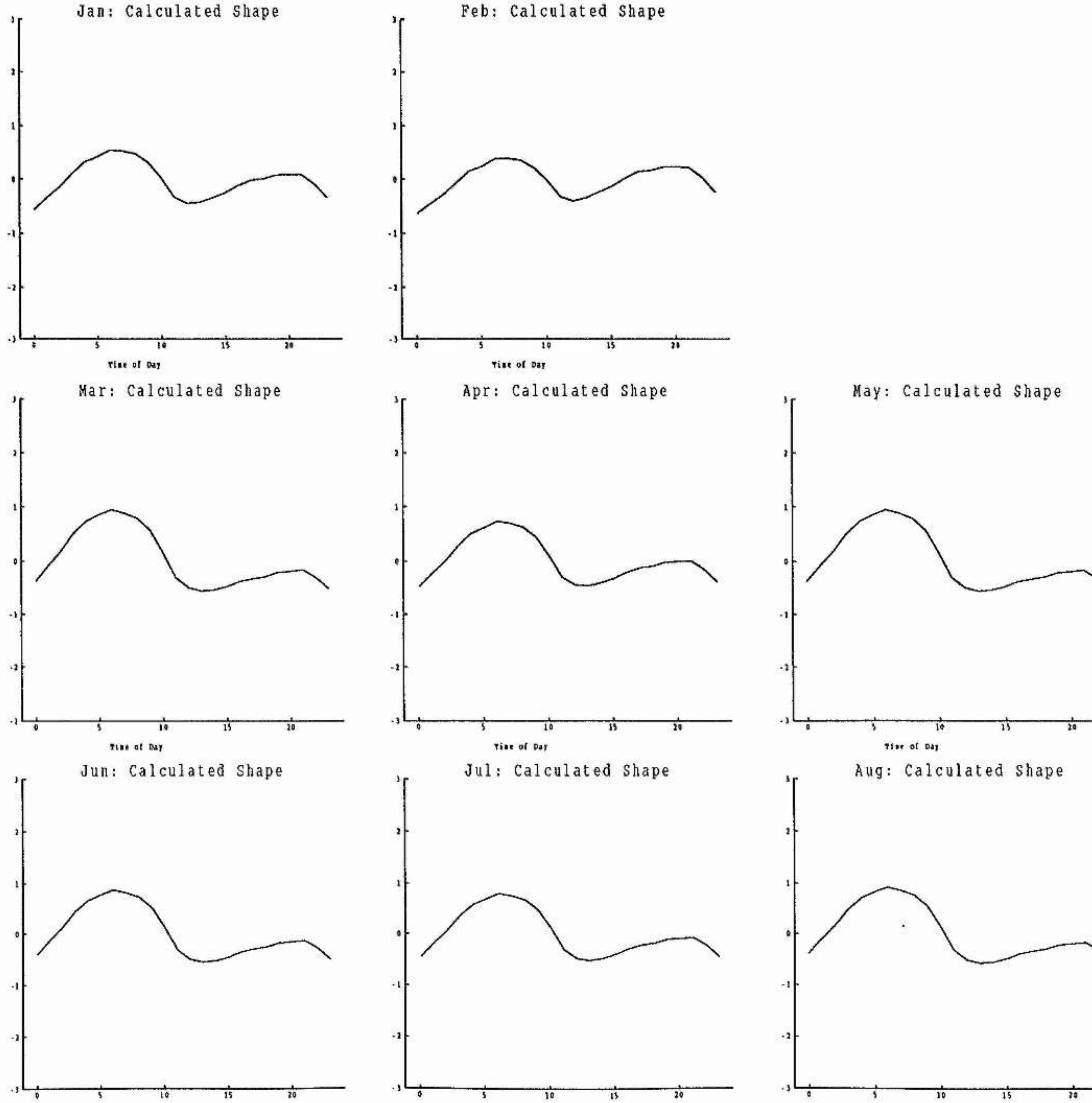
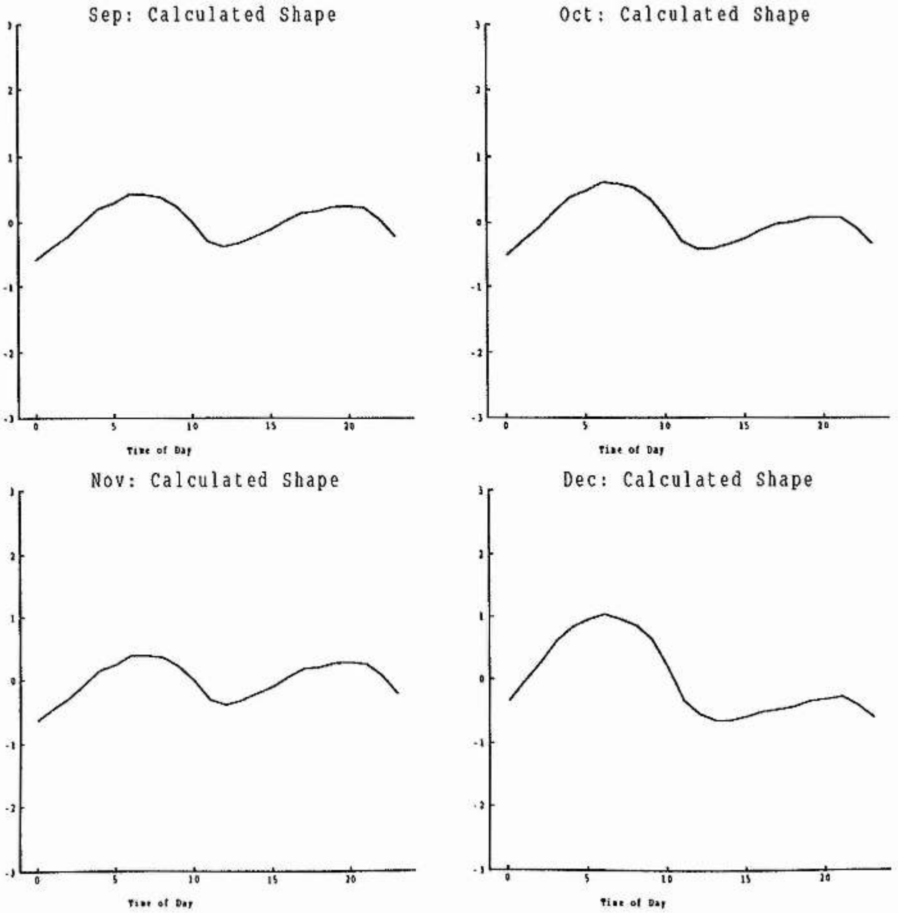


Fig 5.16 (part 4): Weighted average of whole day shapes



D_1 is thought to correspond to days where the rise in ozone levels during the day is due to mixing down of air from higher altitudes which contain a higher ozone concentration than the air at ground level at the end of the night. D_2 is thought to correspond to days where the rise in ozone levels is due to photochemical production, resulting in a maximum about two hours after midday.

Night time shapes

N_1 is not thought to represent any specific physical process, but more a combination of a selection of different days which could not fit into any of the other four clusters. The shape represents only 14 days in 1988, confirming that this shape does not represent any major processes. N_2 and N_4, although superficially quite different, may be variants of the same type of curve, generated by a change in the depletion rates of ozone during the night as the levels of NO_x change. N_3 represents nights during which deposition of ozone occurred at a fairly constant rate, corresponding to a low wind speed during the night.

Mean levels

There is no observable difference between the mean levels of the two types of whole day. This result is unsurprising as it has been shown that the mean ozone level over a day depends primarily on the level over the previous day.

Significant differences were found between the mean levels of the different types of daytime curves however, with those in D_2 being highest. This result is reasonable, as it is days where photochemical production of ozone occurs that would be expected to have the highest observed levels of ozone.

The analysis of the mean levels of the night time shapes showed little difference. This result is as expected as there is no photochemical production of ozone during the night, the only processes governing the night time levels are differing depletion rates.

Variances

Whole days of types W_1 and W_2 showed a marked difference in variance. This result is as expected, as the two whole day types correspond to variation at night and variation during the daytime. As photochemical production occurs during the daytime section of the day, the level of ozone will vary by greater amounts during such days. The results of the analysis of the variances confirm this with days of type two showing markedly greater variances.

The analysis of the variances of the daytime sections shows similar results. The variances of the daytime sections identified with photochemical production and mixing down of ozone are significantly higher.

5.6: Stevenage data set

To investigate whether this method of ozone classification will produce similar results with data from different sites, data have been taken from Stevenage for 1988. Unlike Bush, which is a rural site, Stevenage is an urban site and might be expected to show different results.

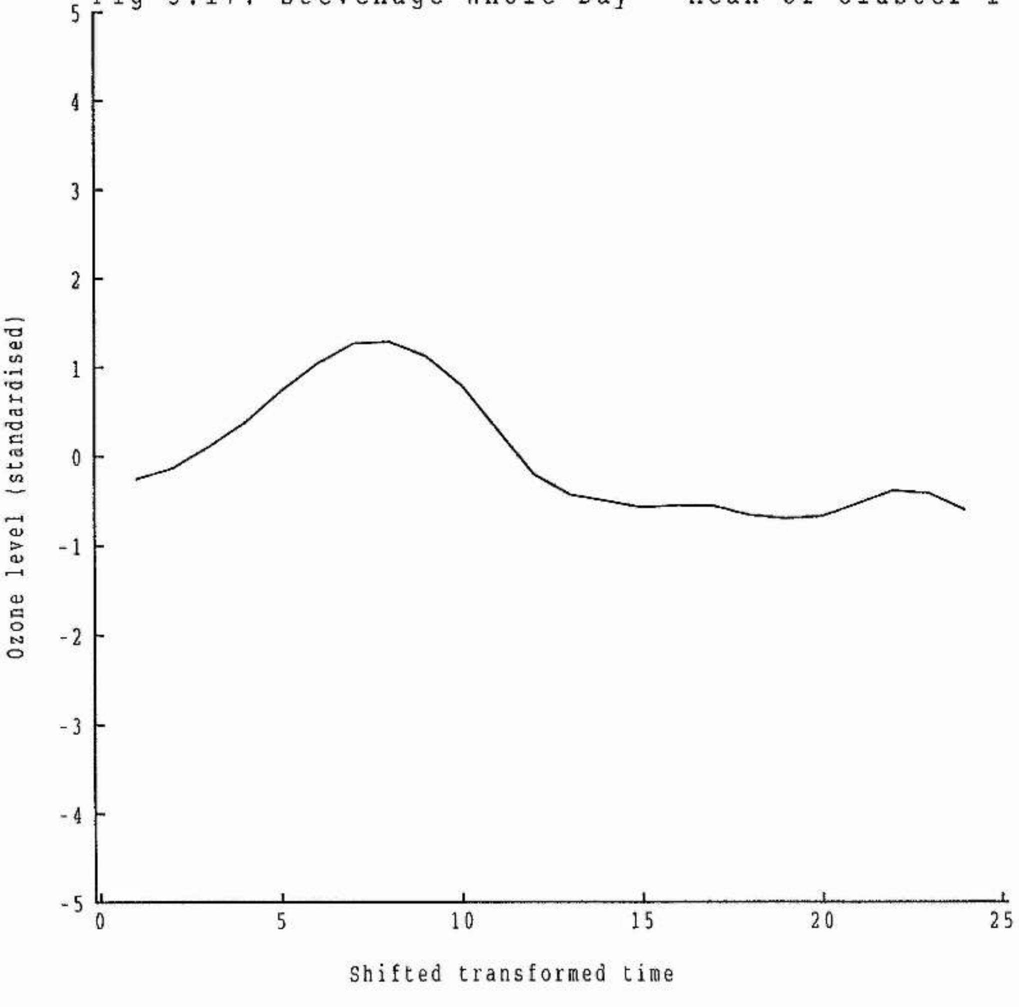
A whole day classification for the data from Stevenage has been carried out in the same way as the analysis of the data from Bush. Data from Stevenage were less complete than those from Bush. After days with missing values were excluded from the analysis, only 119 days were available for clustering.

Table 5.13 summarises the classification results. As at Bush there are two main clusters, one of size 60 and one of size 59, taking the g=2 classification. The two means are shown in Figs 5.17 and 5.18. The two types of day have very similar shapes to those of Bush, and can be interpreted in the same way as before. It should be noted that there are proportionally more days of type 1 for Stevenage. This is as expected for an urban site.

Table 5.13: 1998 Stevenage Whole Day Classification

Number of Clusters	Sizes
10	1,1,4,1,4,2,7,1,45,53
9	1,1,1,4,2,7,1,45,57
8	1,1,1,4,2,1,52,57
7	1,1,1,4,3,52,57
6	1,1,1,4,55,57
5	1,1,4,55,58
4	1,4,56,58
3	4,56,59
2	60,59

Fig 5.17: Stevenage Whole Day - Mean of cluster 1



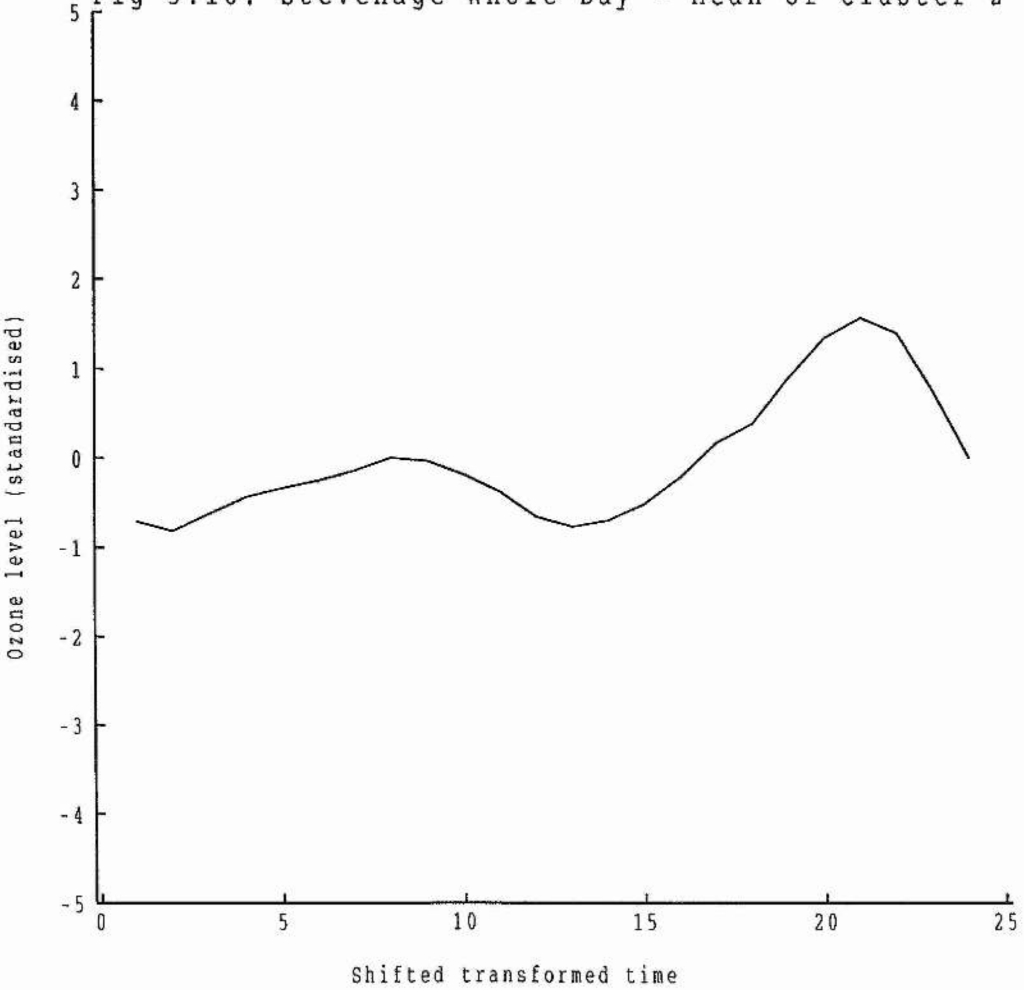


Fig 5.18: Stevenage Whole Day - Mean of cluster 2
5
4
3
2
1
0
-1
-2
-3
-4
-5
Ozone level (standardised)
0
5
10
15
20
25
Shifted transformed time

5.7: Analysis of Weather Factors

To support the interpretations of the different types of whole day, day or night ozone curves given above, data on wind speed, wind direction, solar radiation, rainfall and temperature have been obtained for 1988 at Bush. These data are in the form of averages for each 15 minute period throughout the year.

These 15 minute data have been used to calculate averages for each weather variable for each of the whole day, day or night time sections considered. This has been carried out by calculating the transformed time for each of the 15 minute time points at which a weather observation is available and extracting the sets of observations corresponding to the relevant sections of the transformed time axis. These sets of weather observations have been averaged over each whole day, day or night.

There is a possible problem with calculating average values for wind direction in this manner as the observations are angles in degrees (where 0 corresponds to North). The arithmetic average of two directions of 350 and 10 is 180, whereas the average direction should be 0. The calculation of a circular average is unnecessary in this case as 99% of the wind direction data lies between 80 and 320 and this problem will not adversely affect the analysis overall.

Table 5.14 shows summary statistics of the distributions of the average weather statistics for each of the two types of whole day. The results of pairwise Mann-Whitney tests are also shown. There are significant differences in the distributions of solar radiation, rainfall and temperature. W_2 has on average a higher temperature, less rainfall and more solar radiation than W_1. This supports the interpretations given earlier.

Table 5.14: Whole Day Weather

Weather Variable	Type	Mean	Median	StDev	Mann-Whitney 95% CI and p value for test of equality of medians.
Windspeed	**W1**	2.527	2.299	1.614	(-0.420,0.242)
	W2	2.570	2.066	1.488	p=0.566
Solar Radiation	**W1**	65.83	47.44	61.37	(-63.70,-26.05)
	W2	119.25	102.50	86.79	p=0.000
Rainfall	**W1**	0.034	0.015	0.051	(0.000,0.008)
	W2	0.025	0.005	0.043	p=0.019
Temperature	**W1**	6.855	7.319	3.747	(-2.643,-0.657)
	W2	8.403	8.881	4.233	p=0.002

Table 5.15 shows a comparable summary of the weather for each of the four types of daytime curve. D_1 and D_2 have a higher average solar radiation and temperature. This confirms the interpretations of these days as those corresponding to days where photochemical production of ozone has occurred.

Table 5.15: Daytime Weather

Weather Variable	Type	Mean	Median	StDev
Windspeed	**D1**	2.683	2.016	1.749
	D2	2.759	2.168	1.686
	D3	3.329	3.100	1.835
	D4	2.547	1.923	1.615
Solar Radiation	**D1**	175.10	141.00	122.20
	D2	210.57	204.95	113.56
	D3	126.60	98.10	98.30
	D4	119.70	87.90	110.0
Rainfall	**D1**	0.032	0.000	0.065
	D2	0.023	0.000	0.047
	D3	0.031	0.006	0.057
	D4	0.024	0.005	0.048
Temperature	**D1**	9.392	9.764	4.064
	D2	9.543	10.206	4.403
	D3	6.982	6.359	3.963
	D4	7.594	8.228	4.035

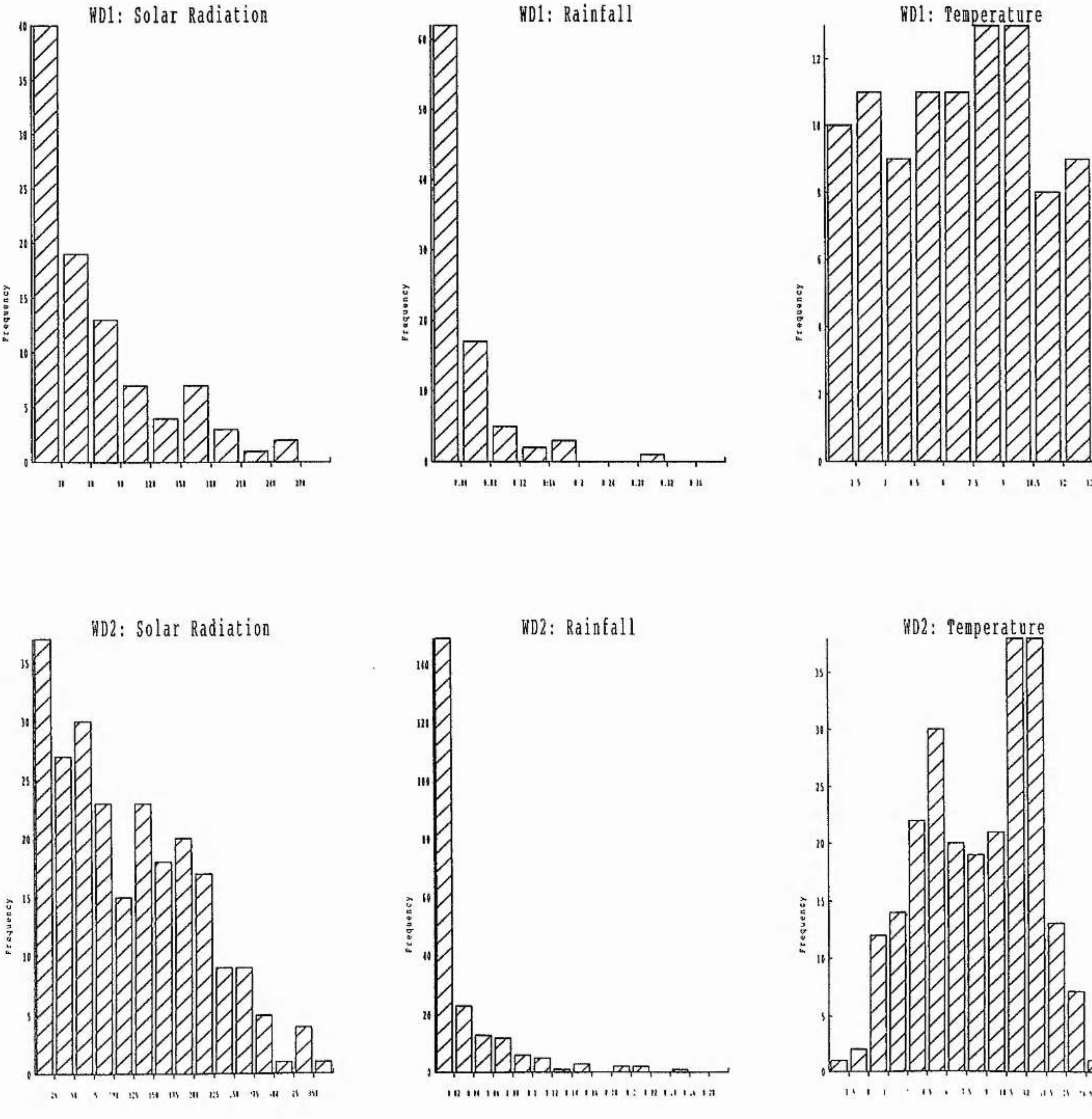
Table 5.16 shows a similar summary of the weather for the five types of night time curve. N1 corresponds to nights with relatively low windspeeds and little rainfall, N2 to relatively high rainfall and low temperatures, N5 to relatively high temperatures and N3 & N4 have very similar weather pattterns to each other. This confirms the hypothesis that the diffent types of curve correspond to different sets of weather conditions overnight.

Table 5.16: Night Time Weather

Weather Variable	Type	Mean	Median	StDev
Windspeed	**N 1**	1.728	1.387	1.220
	N 2	2.317	1.920	1.675
	N 3	2.295	1.680	1.648
	N 4	2.269	1.878	1.653
	N 5	2.433	2.294	1.845
Solar Radiation	**N 1**	2.921	2.121	1.889
	N 2	2.547	1.909	2.120
	N 3	2.961	2.667	2.010
	N 4	3.035	2.500	2.092
	N 5	3.040	2.606	2.087
Rainfall	**N 1**	0.007	0.000	0.010
	N 2	0.138	0.000	0.618
	N 3	0.035	0.000	0.073
	N 4	0.032	0.004	0.079
	N 5	0.018	0.000	0.036
Temperature	**N 1**	6.542	6.744	2.910
	N 2	5.713	5.903	3.899
	N 3	6.931	7.072	3.887
	N 4	6.735	6.714	3.994
	N 5	7.865	8.865	4.464

Fig 5.19 shows histograms of the average solar radiation, rainfall and temperature for the two whole day types. There is a significant difference in location between the distributions

Fig 5.19: Distributions of weather variables for WD1 & WD2



for each type, but there is a considerable overlap. This emphasises the fact that these types of day are not distinct, but each day comes from a continuum of observed patterns.

Table 5.17 summarises a discriminant analysis of the weather averages for the whole day data set. The values of the coefficients of the first CVA vector, compensated to allow for the different scales of the variables by multiplying each by their average, show solar radiation as the most important discriminating variable. This concurs with the observed differences above, as solar radiation shows the most marked difference between the two whole day types. The possibility of rounding errors affecting this analysis due to the different scales of the variables has been checked by carrying out a discriminant analysis on the weather variables adjusted to place them on similar scales. This analysis reached the same conclusion.

Table 5.17: Discriminant Analysis

Variable	CVA1	CVA1 (adj)
Windspeed	-0.178	-0.456
Solar Radiation	-0.013	-2.646
Rainfall	2.637	0.072
Temperature	0.000	0.000

Chapter 6: Discussion of Classification Method

The previous chapter introduced a clustering strategy for ordered data, using an agglomerative algorithm, the 'area between two curves' dissimilarity measure and taking the distance between two clusters to be the distance between the two cluster means. This chapter seeks to place this method in context, by considering its different properties, applying a more standard strategy to the data set already considered and applying the new strategy to two other data sets.

6.1: Properties of Classification Method

Notation

In this chapter the following notation will be used:

d_{ij}	Dissimilarity between two objects i and j.
$d(C_i,C_j)$	Dissimilarity between two clusters C_i and C_j.
n_i	Number of objects in cluster C_i.
x_{it}	Observation on object i, time point t.

Dissimilarity Measure - Metric

As the dissimilarity measure is the area between two curves the following three properties are trivial:

$d_{ij} \geq 0 \qquad \forall i,j \qquad$ **(6.1)**

$d_{ii} = 0 \qquad \forall i \qquad$ **(6.2)**

$d_{ij} = dji \qquad \forall i,j \qquad$ **(6.3)**

The measure can also be shown to satisfy the triangle inequality:

As the total area between two curves is the sum of the areas between sections of the two curves lying between two successive time points, we can obtain properties for the total area by considering small sections of the curves.

As illustrated in Fig 6.1, we consider the points A,B,C on each of three curves i,j,k at one point t on the x axis. As any three points on a straight line satisfy the triangle inequality the following relationship holds:

$$AC \leq AB+BC \qquad \textbf{(6.4)}$$

where

$$AC = |\,(x_{it} - x_{kt})\,|$$

$$AB = |\,(x_{it} - x_{jt})\,|$$

$$BC = |\,(x_{jt} - x_{kt})\,|$$

As the areas between the three curves are the integration of the absolute values of distances which satisfy (6.4) along the x axis then they must also satisfy the triangle inequality.

This result, along with (6.1),(6.2) and (6.3), proves that the area dissimilarity measure is a metric.

Time Ordering of Data

The primary advantage of using the area between two curves dissimilarity measure for objects which consist of ordered data over most standard dissimilarity measures is that the ordering of the observations is considered. With most standard dissimilarity measures (eg. Euclidean) this is not the case. Constructing a measure of dissimilarity can be regarded as an attempt to summarise the data in a meaningful way. When summarising data it is intuitively desirable to make use of as much of the information available as possible, and thus using a measure which takes account of the ordering of the data when such an ordering exists would be preferred to one which does not.

The importance of the time ordering of the data to the area dissimilarity measure is illustrated by the examples displayed in figs 6.2 and 6.3. The curves illustrated differ only in that the y-values at x=2 have been interchanged with those at x=3. Thus these two cases will appear identical to a measure of dissimilarity which does not take account of the ordering of the observations.

In fig 6.2, using the area dissimilarity measure, d_{ij}=1.5; d_{ik}=2.8 and d_{jk}=1.3, leading to the conclusion that j and k are the two most similar objects. In fig 6.3 the area dissimilarity measure gives d_{ij}=1; d_{ik}=2.8 and d_{jk}=1.8, leading to the different conclusion that i and j are the two most similar objects. Thus an agglomerative classification algorithm would reach a different conclusion in each of these examples as to which two objects to join at this stage.

However, using Euclidean distance, d_{ij}=1.4; d_{ik}=2.4 and d_{jk}=1.5 for both examples, leading to the decision to join i and j.

Figs 6.1 - 6.3: Area dissimilarity measure

Fig 6.1

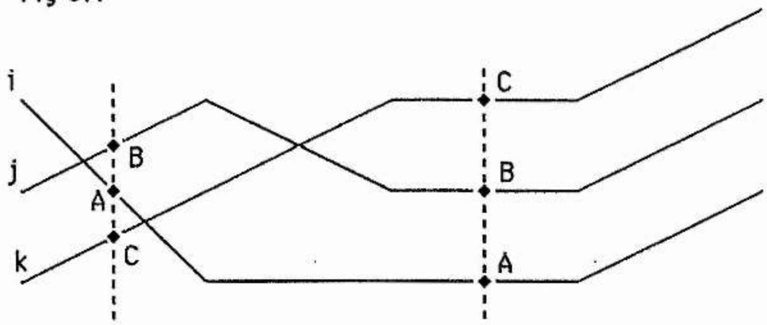


Fig 6.2

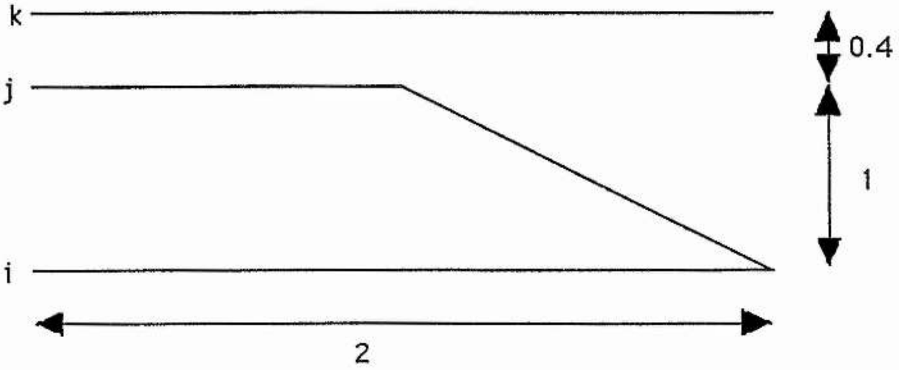
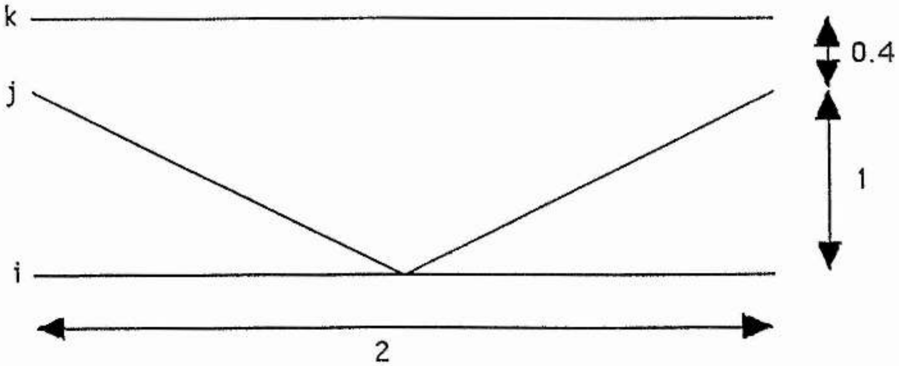


Fig 6.3



As a further example, consider the dissimilarity measure defined as:

$$d_{ij} = \sum_{t} \mid (x_{it} - x_{jt}) \mid$$

Using this dissimilarity measure, for both of the examples considered above, $d_{ij}=2$; $d_{ik}=4.2$ and $d_{jk}=2.2$, leading to the conclusion to join i and j.

This example demonstrates the importance of the time ordering of the data to the area dissimilarity measure as a change in the ordering has resulted in different values for the dissimilarities and a different decision for which two objects to join, whilst the same change in ordering has had no effect on the values taken by two standard dissimilarity measures. This lack of consideration of the ordering would also be true for all of the most commonly used measures.

Space Contracting/Dilating

Lance and Williams (1967) introduced the concept of space distortion. A clustering strategy is said to be space-contracting if a new cluster, on formation, will appear to move nearer to some or all of the remaining clusters (i.e. $d(C_i,C_k) \geq d(C_i \cup C_j , C_k)$) and space-dilating if it appears to recede (i.e. $d(C_i \cup C_j , C_k) \geq d(C_j,C_k)$). This strategy can behave both as space contracting or dilating at different stages. Examples of each of these are shown in figs 6.4 and 6.5 respectively.

As this clustering strategy can be space-contracting it does have the controversial property that it can produce reversals in the classification, i.e. that $d(C_i,C_j) > d(C_i \cup C_j , C_k)$. An example of this is displayed in fig 6.6, where $d(C_i,C_j)=2.4$; $d(C_i,C_k)=2.5$; $d(C_j,C_k)=2.7$ and $d(C_i \cup C_j , C_k)=2.1$.

Figs 6.4 - 6.6: Space-dilation property of algorithm

Fig 6.4

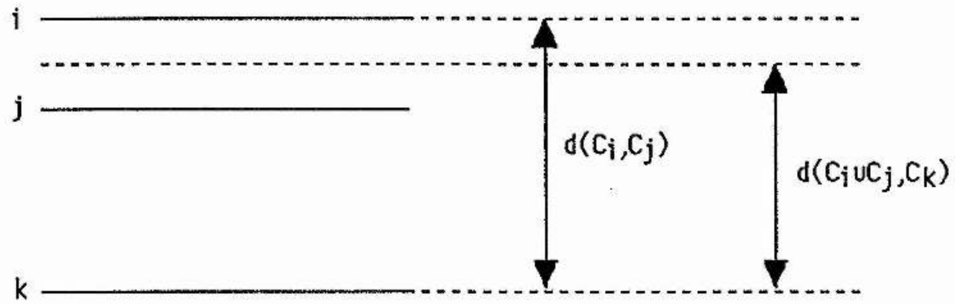


Fig 6.5

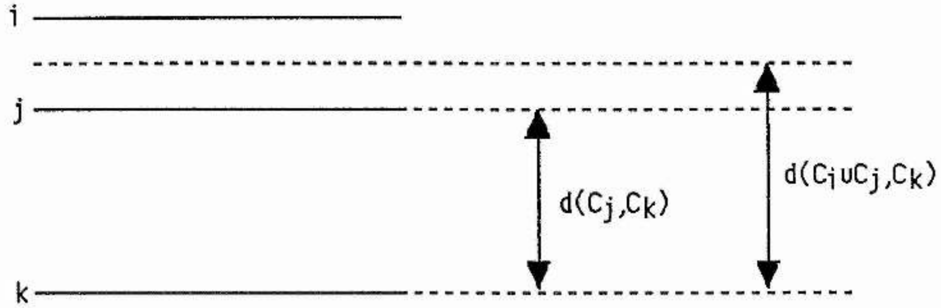
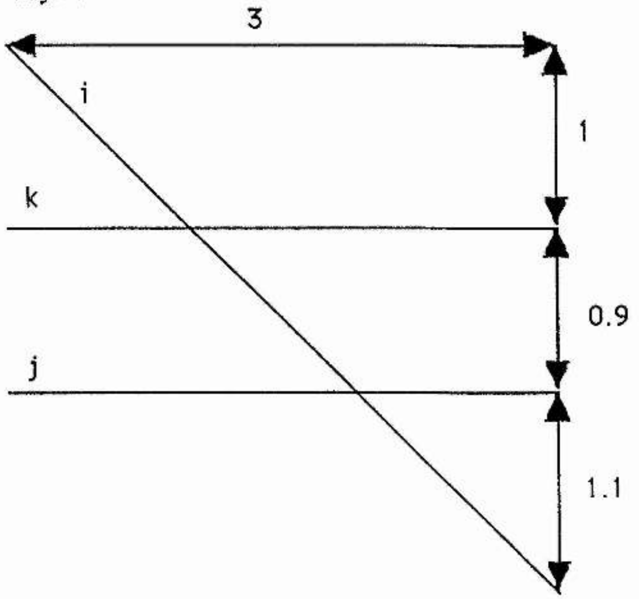


Fig 6.6



Non-Combinatorial

Suppose that two clusters C_i and C_j are to be joined to form C_h. For many dissimilarity measures $d(C_k,C_h)$ can be expressed as a linear function of $d(C_i,C_k)$, $d(C_j,C_k)$, $d(C_i,C_j)$, n_i and n_j. Lance and Williams (1967) defined strategies with this property as combinatorial.

This strategy is non-combinatorial. To see this, examine the examples illustrated in figs 6.7, 6.8, 6.9 and 6.10.

If we assume that a relationship of the form:

$$d(C_k,C_h) = \alpha_1\, d(C_i,C_j) + \alpha_2\, d(C_j,C_k) + \alpha_3\, d(C_i,C_k) + \text{const} \qquad \mathbf{(6.5)}$$

exists. Then figs 6.7 and 6.8 imply that $\alpha_1 = 0$ as only $d(C_i,C_j)$ differs between them. Similarly figs 6.7 and 6.9 imply that $\alpha_2 = 0$. If we then consider figs 6.7 and 6.10 and set α_1 and α_2 equal to zero then 6.7 implies $\alpha_3 = 3/3.5$ and 6.10 implies $\alpha_3 = 1$. Thus (6.5) cannot hold.

The advantage of using a combinatorial strategy is that the data may be kept as distance matrices during the clustering procedure which increases the speed of computation. In this case the data are kept as curves and the computing time was found to be in the region of 1 hour to complete the procedure on the ozone data set consisting of 336 objects of 24 variables considered in the previous chapter.

Reducibility

Gordon (1987) mentions the concepts of reducibility and reciprocal nearest neighbours. If a clustering strategy satisfies the condition

$$d(C_i \cup C_j , C_k) \geq \min (d(C_i,C_k) , d(C_j,C_k)) \qquad \forall C_k \qquad \textbf{(6.6)}$$

then the computing time requirements can be reduced by joining pairs of reciprocal nearest neighbours in a single pass.

The example displayed earlier in fig 6.6 provides an example where condition (6.6) does not hold, so this is not possible in this case.

Figs 6.7 - 6.10: Non-combinatorial property of algorithm

Fig 6.7

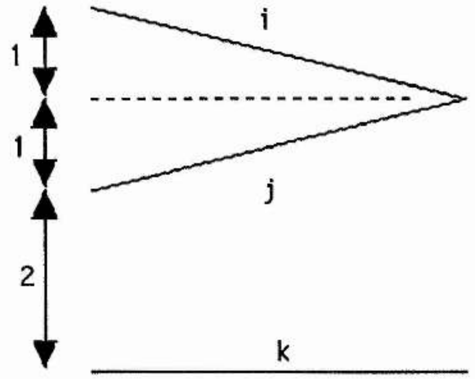


Fig 6.8

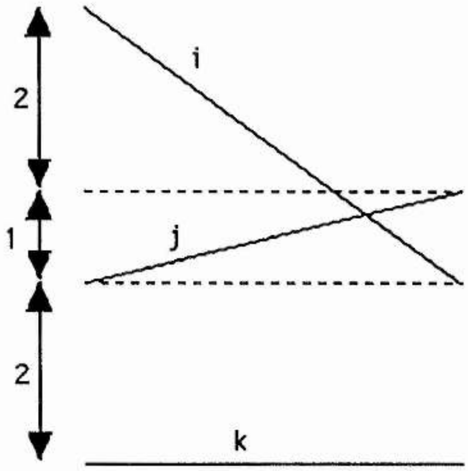


Fig 6.9

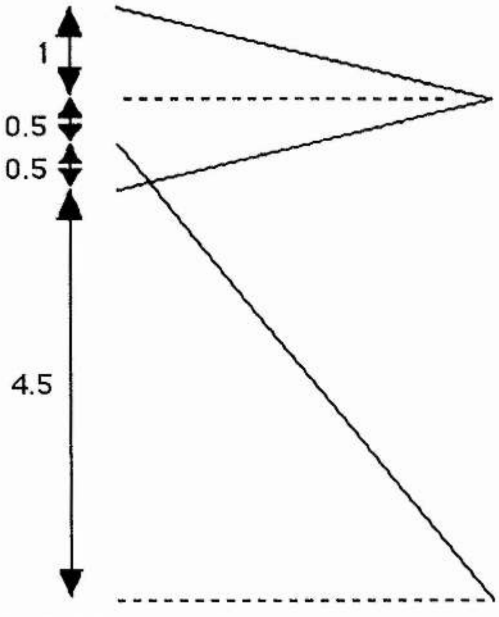
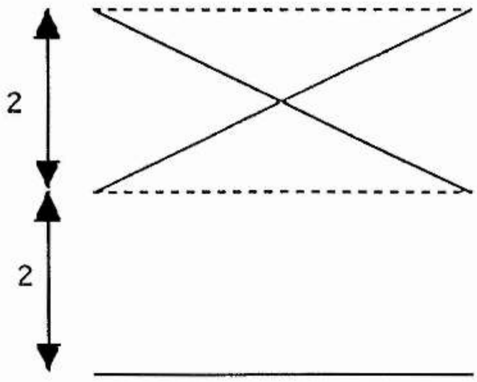


Fig 6.10



Relationship to Group Average Link

The difference between two clusters C_i and C_j has been defined as the difference between the two cluster means (MC method). As was indicated in chapter 5 with many standard dissimilarity measures (e.g. Euclidean distance) this would be equivalent to taking the mean of all the pairwise differences between objects in C_i and C_j, commonly referred to as the group average link method (AL method).

If none of the curves in C_i and C_j cross then these methods are equivalent as the examples in figs 6.7 and 6.9 demonstrate. In fig 6.7

$$d(C_k,C_h) = 0.5\ d(C_j,C_k) + 0.5\ d(C_i,C_k) \qquad \textbf{(6.7)}$$

but in fig 6.9 k crosses j and (6.7) does not hold.

Thus if there are no crossings between any of the curves in the data at any stage of the algorithm, the MC and AL methods will be equivalent. Conversely, if there are crossings then the results obtained from the two methods will differ.

6.2: Classification of Ozone Data using AL Method

In order to investigate the difference between the MC and AL methods further, a classification of the Bush ozone data used in chapter 5 has been carried out using the AL method. Tables 6.1, 6.2 and 6.3 summarise the classifications thus obtained for the whole day, daytime only and night time only data sets, for g=10 to g=2 where g represents the number of clusters that exist at the stage of the classification algorithm being considered. These can be compared to tables 5.1, 5.2 and 5.3 from chapter 5.

Table 6.1: 1988 Bush AL Classification : Whole Day

Number of Clusters	Sizes
10	2,2,5,1,23,23,20,13,59,188
9	2,5,1,23,23,22,13,59,188
8	5,1,23,23,22,13,59,190
7	5,23,24,22,13,59,190
6	5,23,22,13,83,190
5	5,23,22,83,203
4	28,22,83,203
3	28,105,203
2	133,203

Table 6.2: 1988 Bush AL Classification : Daytime Only

Number of Clusters	Sizes
10	6,3,12,16,12,25,78,148,27,9
9	6,3,12,10,25,78,160,27,9
8	6,3,16,256,78,172,27,9
7	6,3,41,78,172,27,9
6	6,3,41,172,105,9
5	9,41,172,105,9
4	9,172,105,50
3	9,277,50
2	9,327

Table 6.3: 1988 Bush AL Classification : Night Time Only

Number of Clusters	Sizes
10	6,7,11,10,9,57,66,39,113,18
9	6,7,11,9,57,66,49,113,18
8	6,11,9,57,66,49,113,25
7	6,11,9,66,106,113,25
6	6,11,75,106,113,25
5	6,11,106,188,25
4	11,112,188,25
3	112,199,25
2	199,137

Tables 5.1 and 6.1 indicate that both methods are in agreement in that they have discovered two main clusters in the whole day data, although this is much less pronounced in the AL results as more days have been allocated to the smaller clusters.

Figs 6.11 and 6.12 display the means of the two clusters obtained at the g=2 stage of the AL classification. These two mean curves are very similar to the two shapes obtained previously by the MC method (figs 5.2 and 5.3).

Figs 6.13 displays a principal co-ordinate plot of the g=4 AL classification which can be compared to fig 5.13 from chapter 5. The two main clusters are similar, and the extra clusters formed by the AL method exist in the boundary between them.

Tables 5.2 and 6.2 indicate that with the AL classification the choice of clusters for the daytime data is now even harder than before. The AL method has again increased the size of the smaller clusters observed with the MC method, to the extent that a choice of a small number of clusters is very difficult. This pattern is repeated with the results from the night time data, as displayed in tables 5.3 and 6.3.

These results lead to the conclusion that the MC method tends to produce a smaller number of larger clusters, whereas the AL method tends to produce a larger number of smaller clusters.

Fig 6.11: Mean of cluster 1 (g=2, AL method)

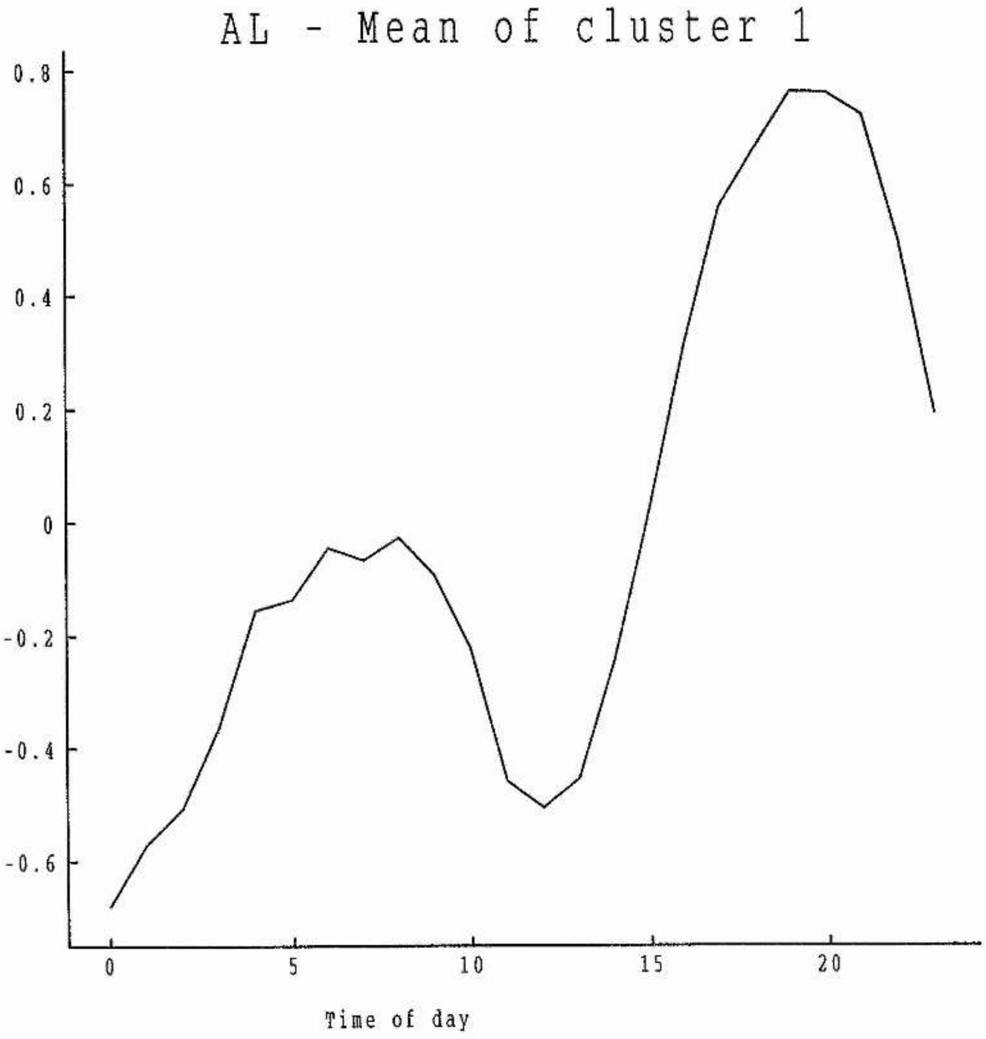


Fig 6.12: Mean of cluster 2 (g=2, AL method)

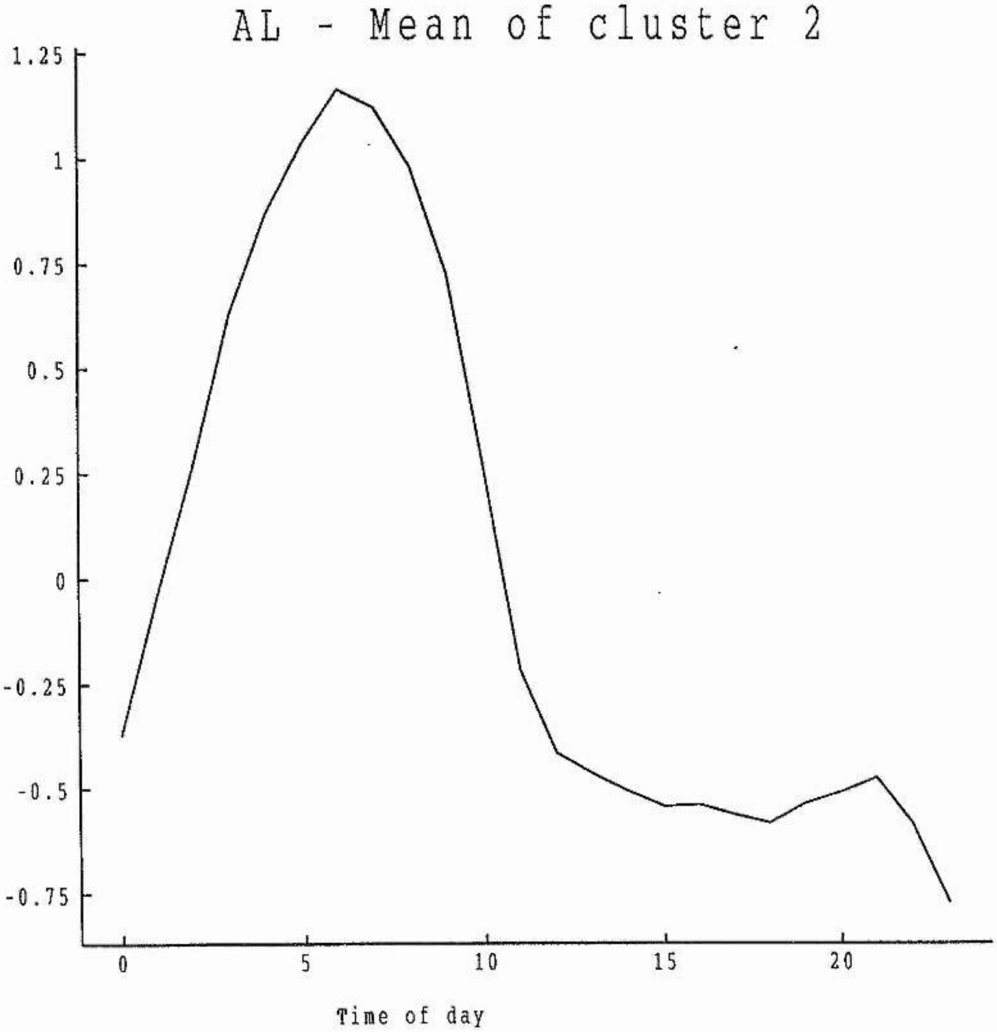
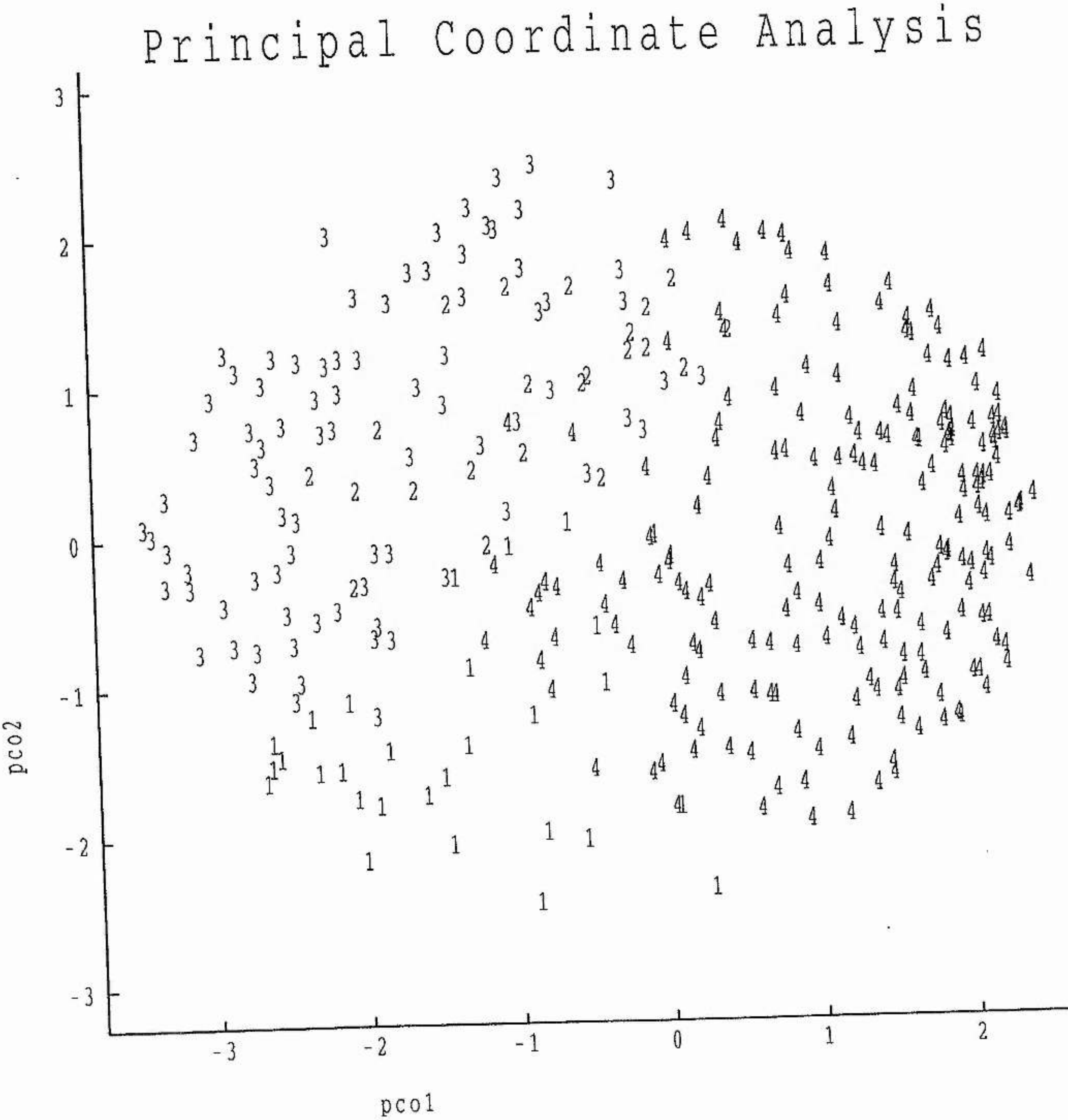


Fig 6.13: Principal Co-ordinate Plot (g=4, AL method)



6.3: Application to Stock Yield Data

The MC classification method has been applied to the data contained in table 11.13 of Hartigan (1975). The data are yield of stocks for 34 companies (objects) over the years 1959-1969.

Figs 6.14 and 6.15 display the results of a principal co-ordinate analysis of these data, with each object represented by a number 1-34. These suggest the existence of about 3-5 quite separate clusters.

Table 6.4 summarises the classifications from g=10 to g=4. An examination of table 6.4 also suggests 3-5 main clusters. Figs 6.16 and 6.17 display the same principal co-ordinate analysis with each point represented by a number from 1-5 indicating which of the 5 clusters that object has been allocated to from the g=5 classification. Clusters 2,3 and 5 are separate on pco 1 and 2, and clusters 1 and 4 separate on pco 3.

Table 6.4: Stock yield data MC classification

Number of clusters	Sizes
10	1,1,2,4,1,5,9,1,3,7
9	1,1,4,3,5,9,1,3,7
8	1,1,4,3,9,6,3,7
7	1,4,3,9,6,4,7
6	1,3,9,6,4,11
5	3,9,6,4,12
4	3,6,13,12

This data set has also been classified using the Euclidean dissimilarity metric and the average linkage (ALE, equivalent to dissimilarity between cluster means in this case), single linkage (SLE, nearest neighbour) and complete linkage (CLE, farthest neighbour) methods. The results from the five cluster stage of the ALE method are displayed as before in figs 6.18 and 6.19. The clustering is identical to that obtained using the MC method, except that objects 21,29 and 19 have moved between clusters 2 and 3. There is nothing exceptional

Fig 6.14: Principal Co-ordinate Plot: Stock Yield Data

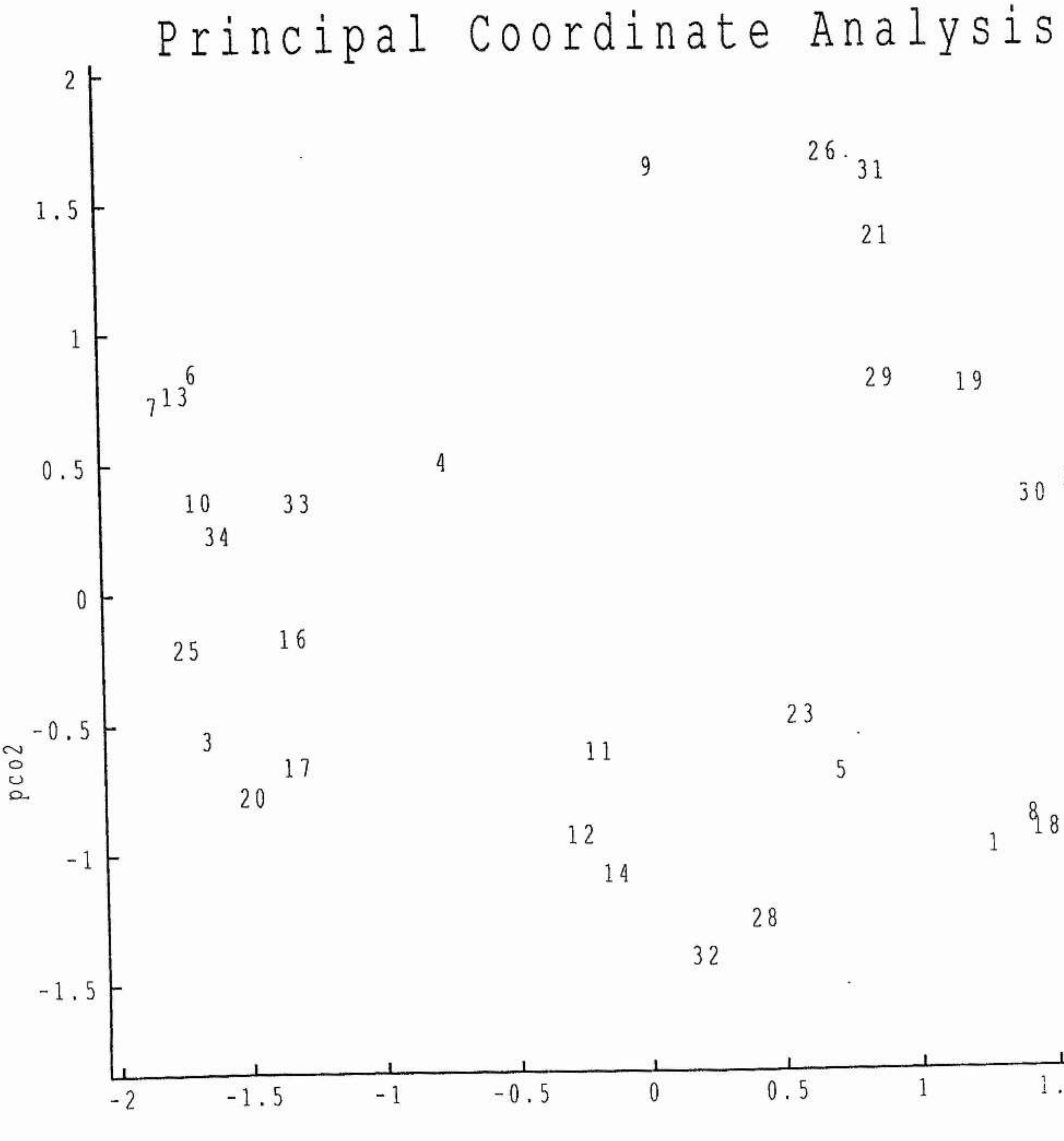


Fig 6.15: Principal Co-ordinate Plot: Stock Yield Data

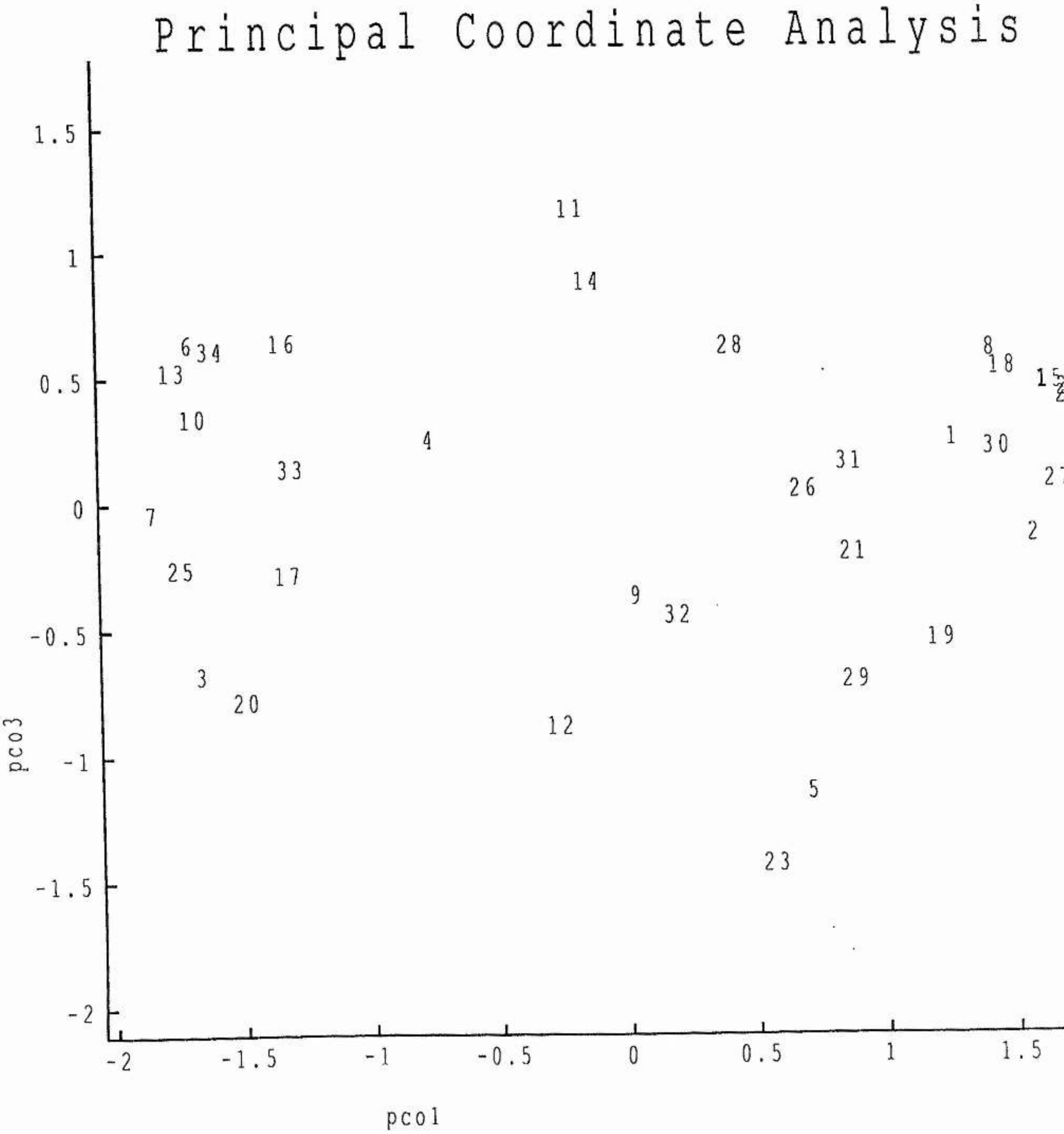


Fig 6.16: Principal Co-ordinate Plot: Stock Yield Data (MC method, g=5)

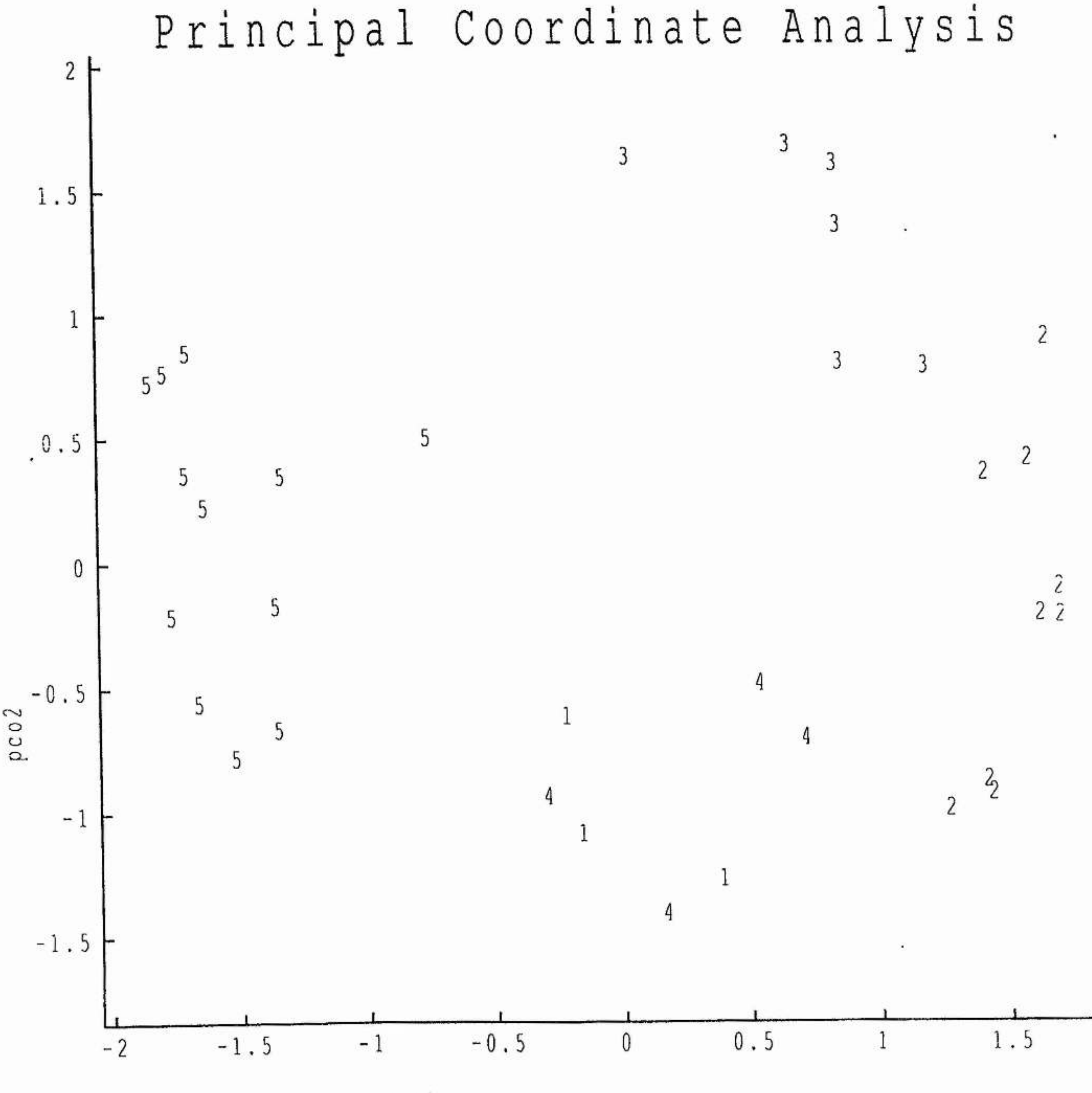


Fig 6.17: Principal Co-ordinate Plot: Stock Yield Data (MC method, g=5)

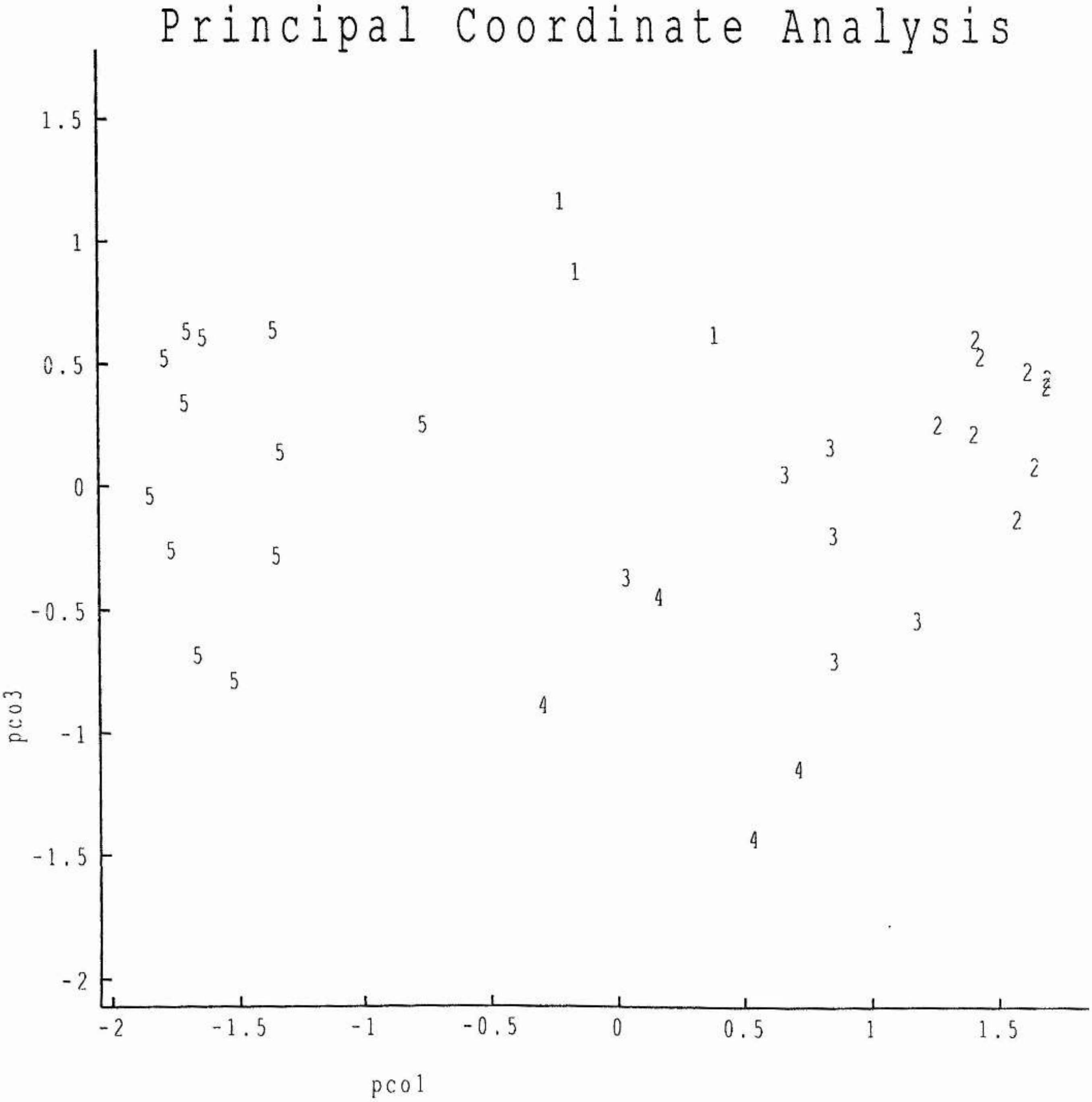


Fig 6.18: Principal Co-ordinate Plot: Stock Yield Data (ALE method, g=5)

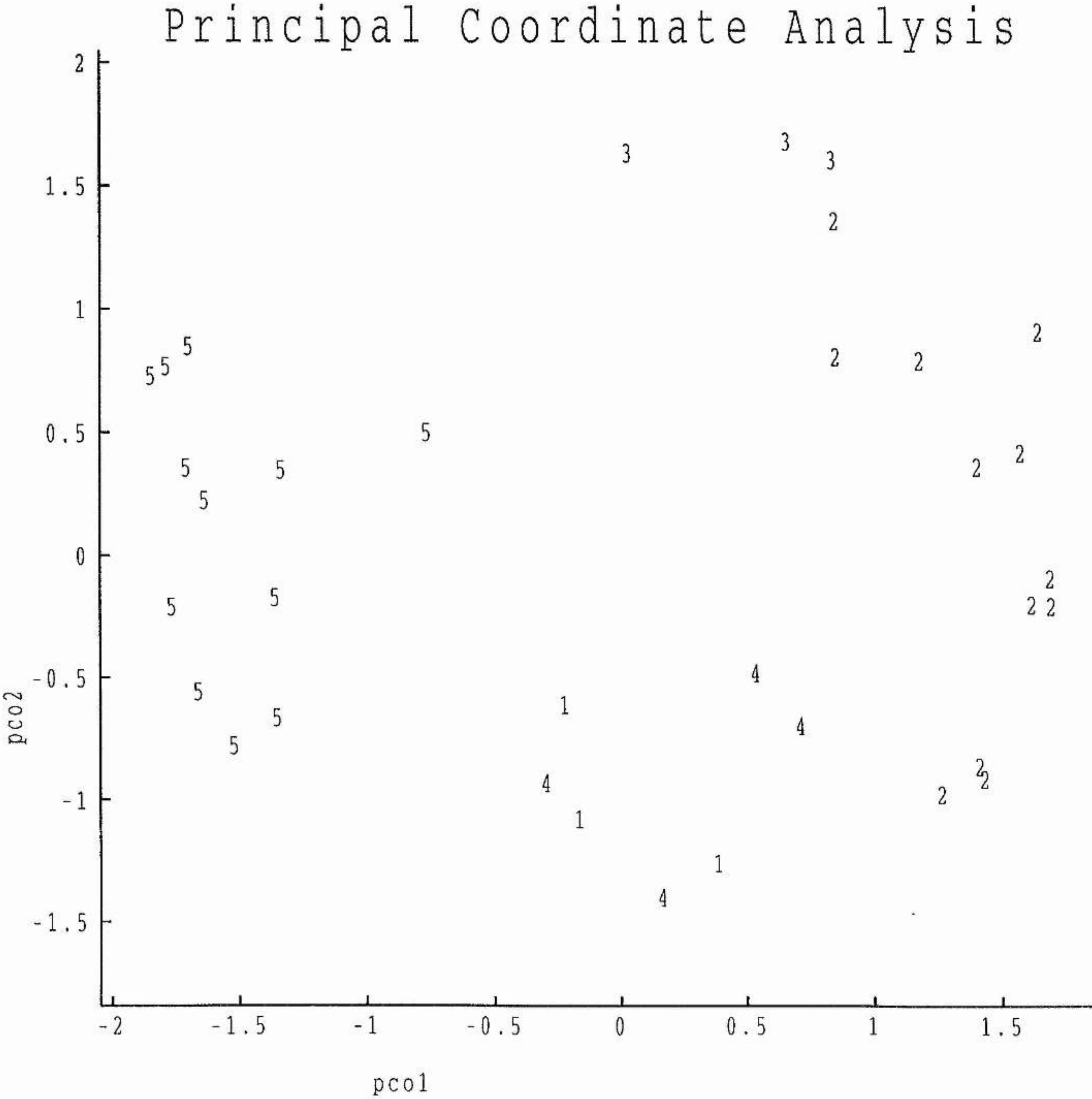
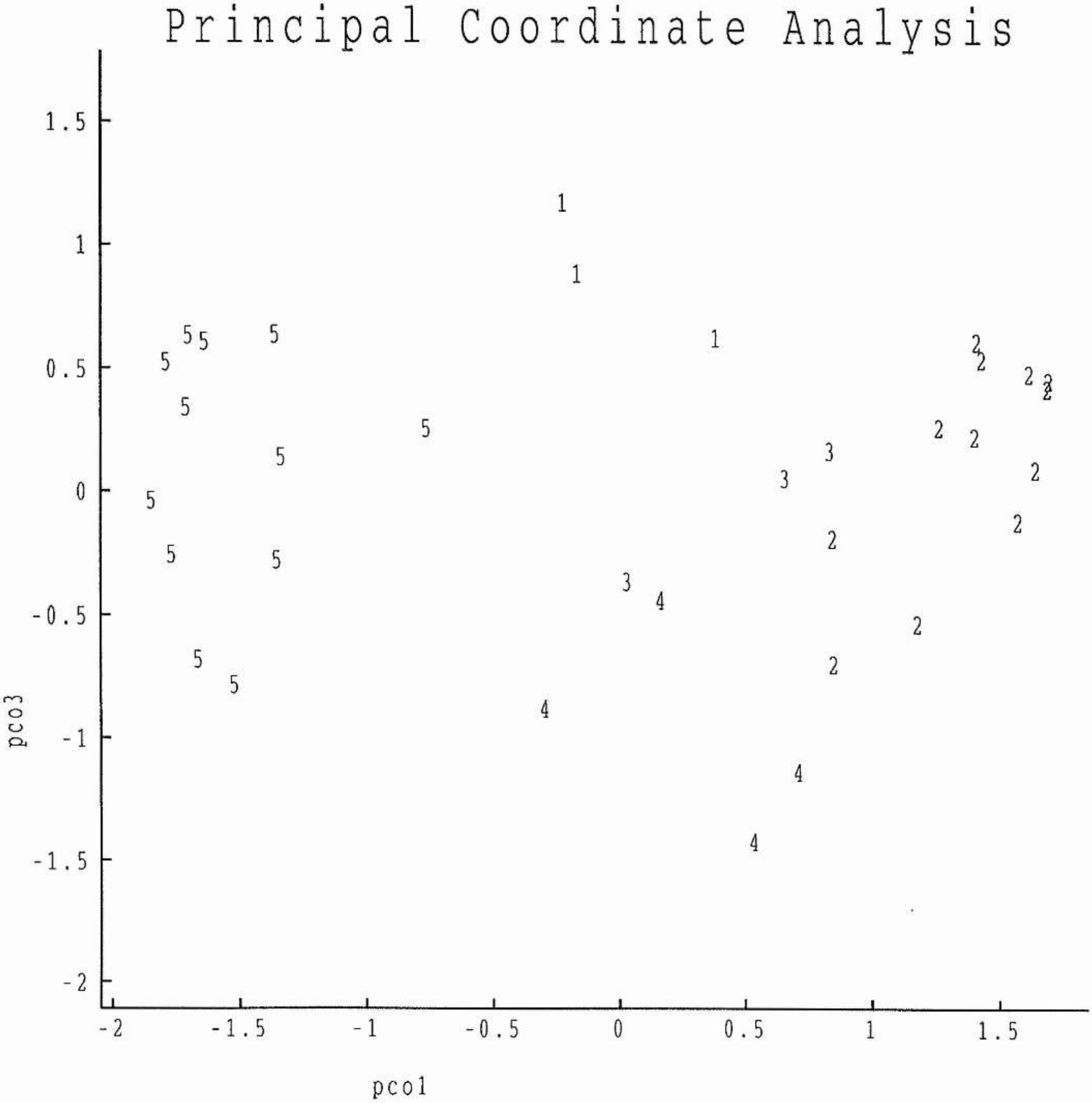


Fig 6.19: Principal Co-ordinate Plot: Stock Yield Data (ALE method, g=5)



about these 3 objects, but they do lie between the centres of clusters 2 and 3 and could reasonably be placed in either.

Figs 6.20 and 6.21 display the results obtained from the five cluster stage of the SLE method. The 'chaining' properties for which the single linkage method is notorious are evident. There are only three clusters out of the five which contain more than one object. Cluster 3 corresponds to the bulk of cluster 5 (MC), cluster 5 corresponds to clusters 2,3 and 4 (MC), cluster 4 corresponds to cluster 1 (MC) and of the two single objects, only cluster 1 (object 4) can be regarded as lying outside the main clusters.

Figs 6.22 and 6.23 display the results obtained from the five cluster stage of the CLE method. The space dilating properties of this method have caused the original cluster 5 to be split up into 2 separate clusters, while clusters 1 and 2 have joined together.

The comparison between the MC and ALE methods follows the similarity between the results obtained from the AL and MC methods above, namely that the bulk of the classification is very similar, with objects lying between two cluster centres liable to be classified differently. The comparison with the SLE and CLE methods demonstrates that the MC method does not share their undesirable space-distorting properties.

Fig 6.20: Principal Co-ordinate Plot: Stock Yield Data (SLE method, g=5)

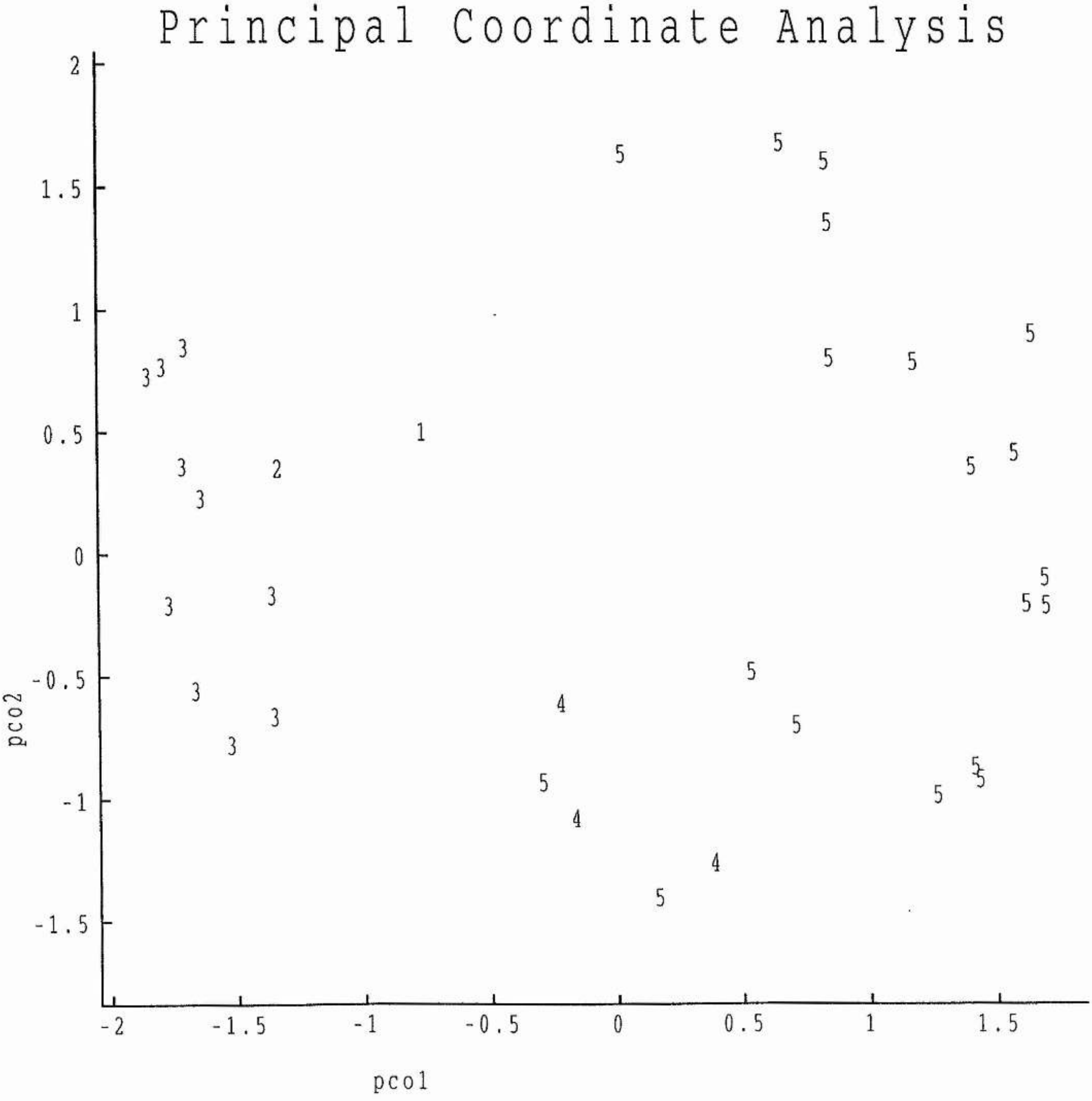


Fig 6.21: Principal Co-ordinate Plot: Stock Yield Data (SLE method, g=5)

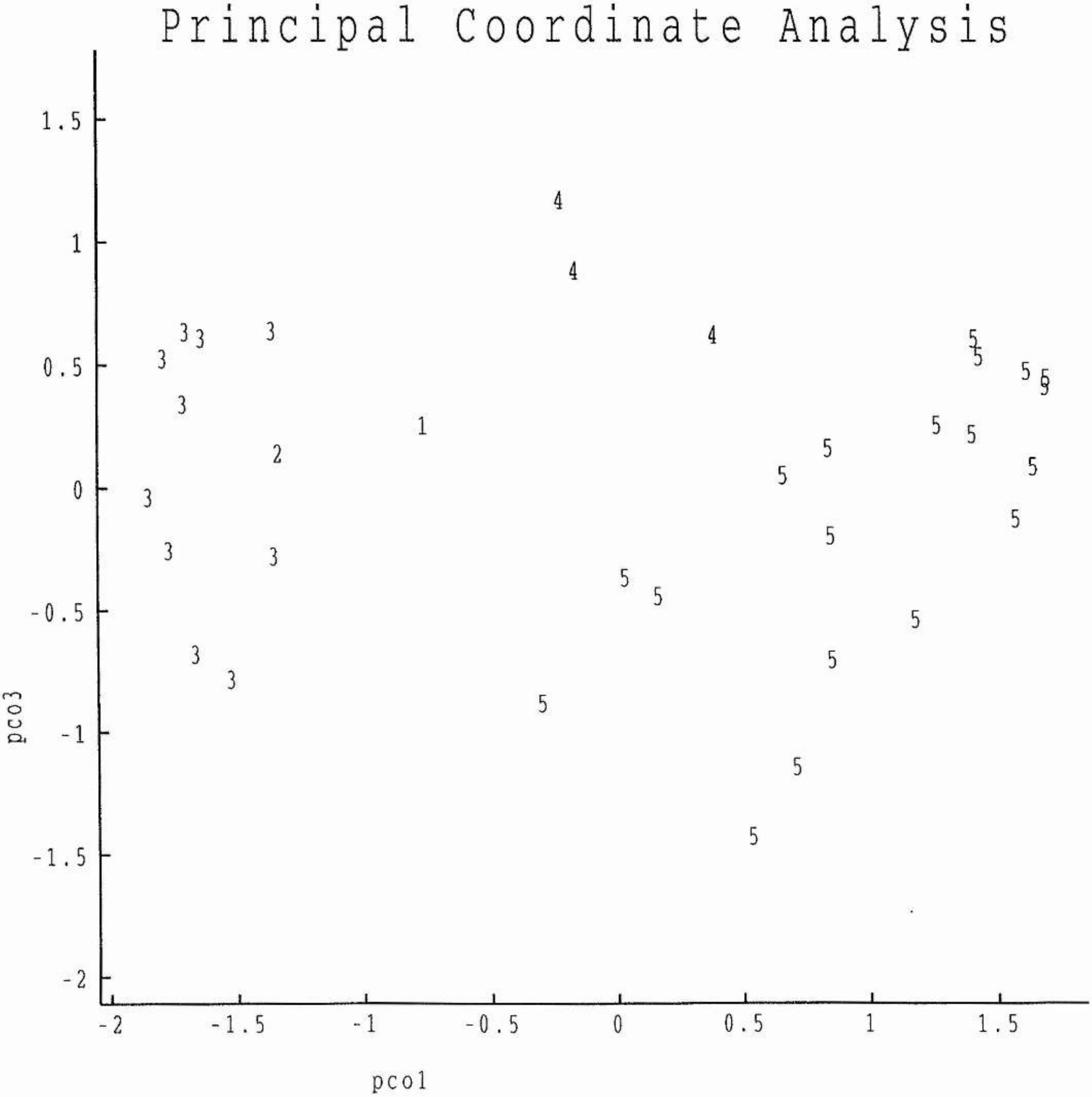


Fig 6.22: Principal Co-ordinate Plot: Stock Yield Data (CLE method, g=5)

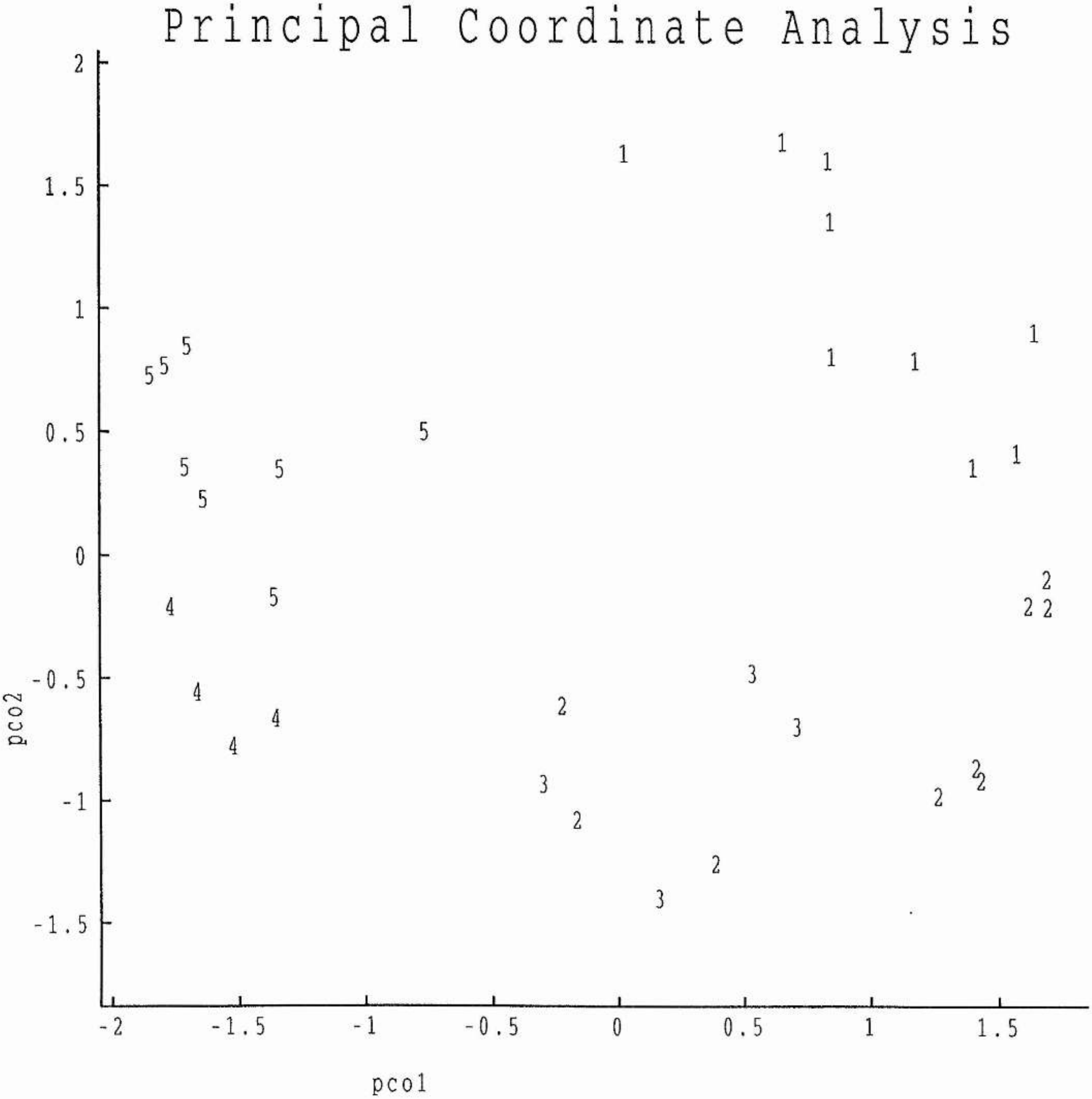
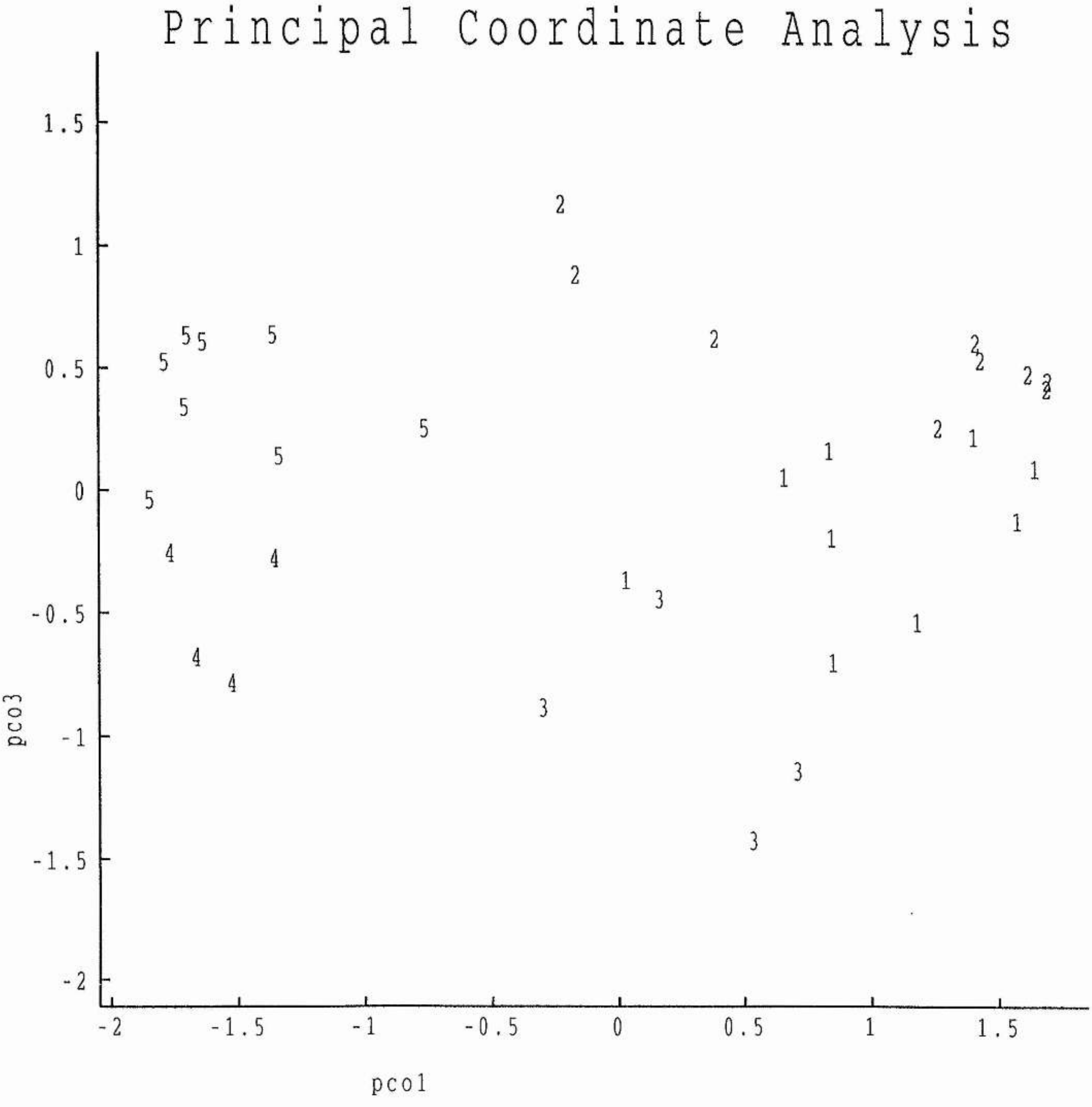


Fig 6.23: Principal Co-ordinate Plot: Stock Yield Data
(CLE method, g=5)



6.4: Principal Points of Ozone Data

Flury (1990) introduced the idea of principal points of a given distribution. These principal points are the optimal cluster means of a k-means clustering procedure and are defined as a set of k points in p dimensional space that optimally approximate a given distribution in terms of least squares. The problem addressed by principal points is to find optimal estimates of cluster means, for a given number of clusters, whereas classification techniques are more concerned with finding the clusters themselves. The work carried out in chapter 5 on the Bush ozone data indicated that there are two main types of whole day curve and gave estimates of the mean curve for each type. Thus it will be of interest to obtain estimates of the two principal points of the distribution of these data, for comparison with the cluster means already obtained.

Flury (1993) proposed estimating the principal points of a distribution by carrying out a k-means clustering algorithm on the values of the first q principal components.

Table 6.5 summarises a principal components analysis of the whole day ozone data from chapter 5. The data have already been standardised, so the covariance matrix was used. The first 12 principal components account for 94.3% of the total variation and have been chosen to carry out the estimation as the last 12 components make very small contributions to the total variance explained.

Table 6.5: Principal components analysis of whole day data

PC	1	2	3	4	5	6	7	8	9
% Var	35.19	17.25	11.98	8.62	5.54	4.39	2.91	2.63	1.93
PC	10	11	12	13	14	15	16	17	18
% Var	1.54	1.33	1.01	0.89	0.82	0.80	0.60	0.52	0.48
PC	19	20	21	22	23	24			
% Var	0.42	0.35	0.28	0.26	0.24	0.00			

Fig 6.24 shows the position of the 2 principal points (PP1 & PP2) obtained by a k-means procedure on the first 12 principal component scores and the cluster means obtained in chapter 5 (W_1 and W_2) on a plot of the first 2 principal component scores. The principal points lie very close to these cluster means.

Fig 6.25 shows the ozone curves corresponding to PP1 and PP2 alongside those from W_1 and W_2. The two sets of curves do follow similar patterns, although those from the principal points appear to be simpler. This is due to the fact that the curves from the principal points have been generated using the first 12 principal components and thus some of the detail has been lost.

The work carried out in this section has provided further evidence of the validity of the MC approach in the case of the Bush ozone data, as it has found similar estimates of the two types of 'typical' day.

Fig 6.24: Position of Principal Points (PP1 & PP2): Ozone Data

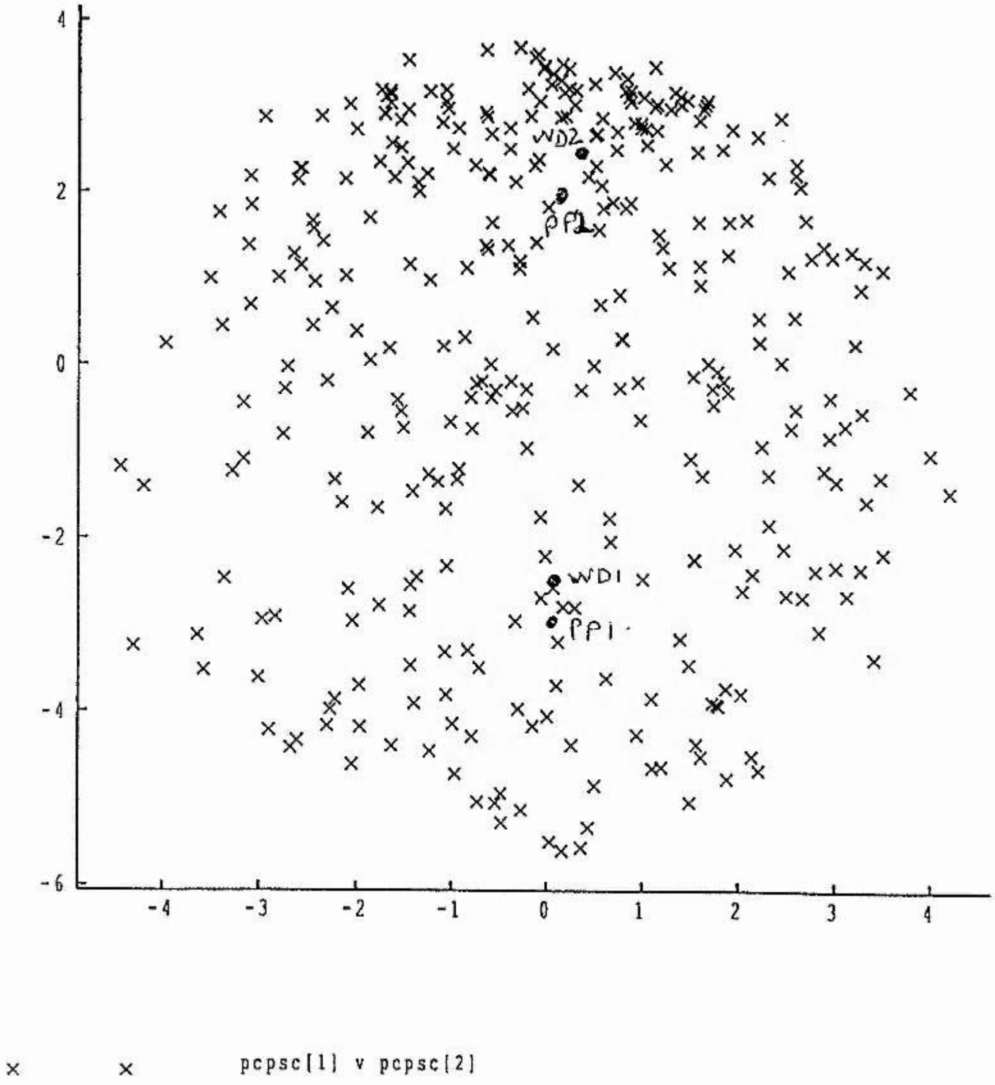


Fig 6.25 (Part 1): Ozone Curve: Principal Point PP1

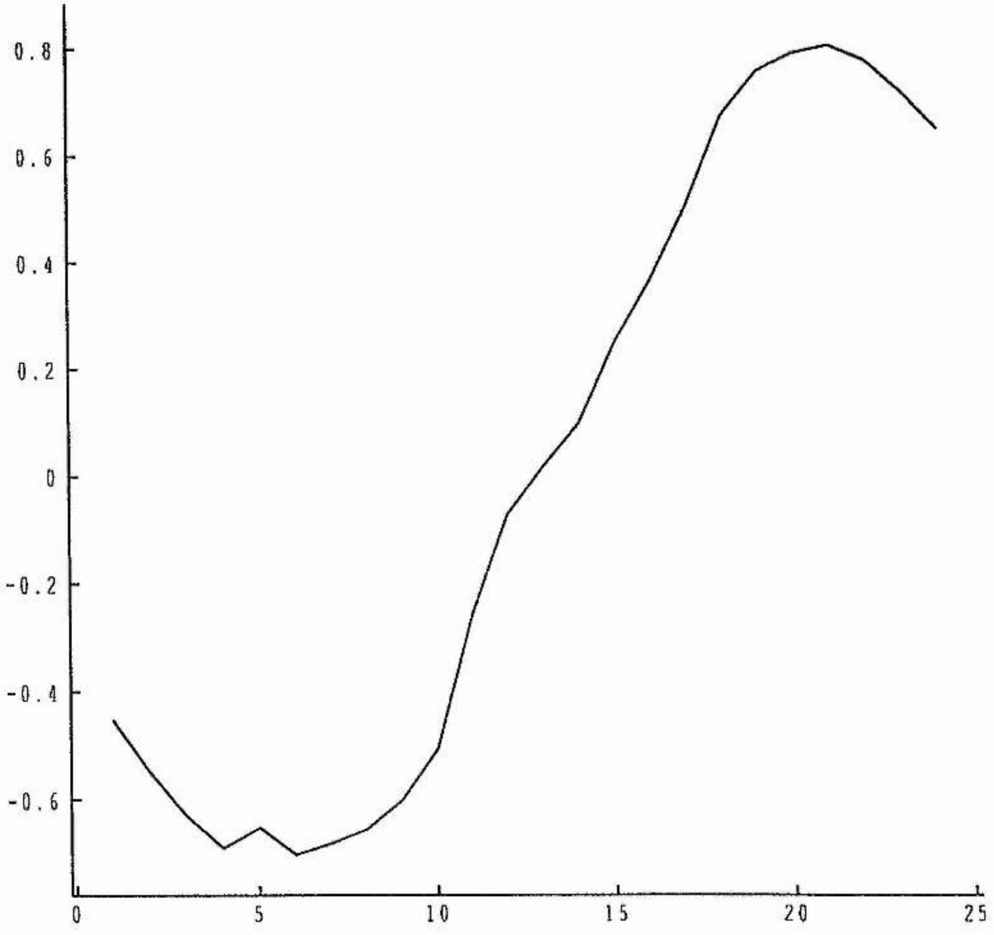


Fig 6.25 (Part 2): Ozone Curve: Whole Day WD2

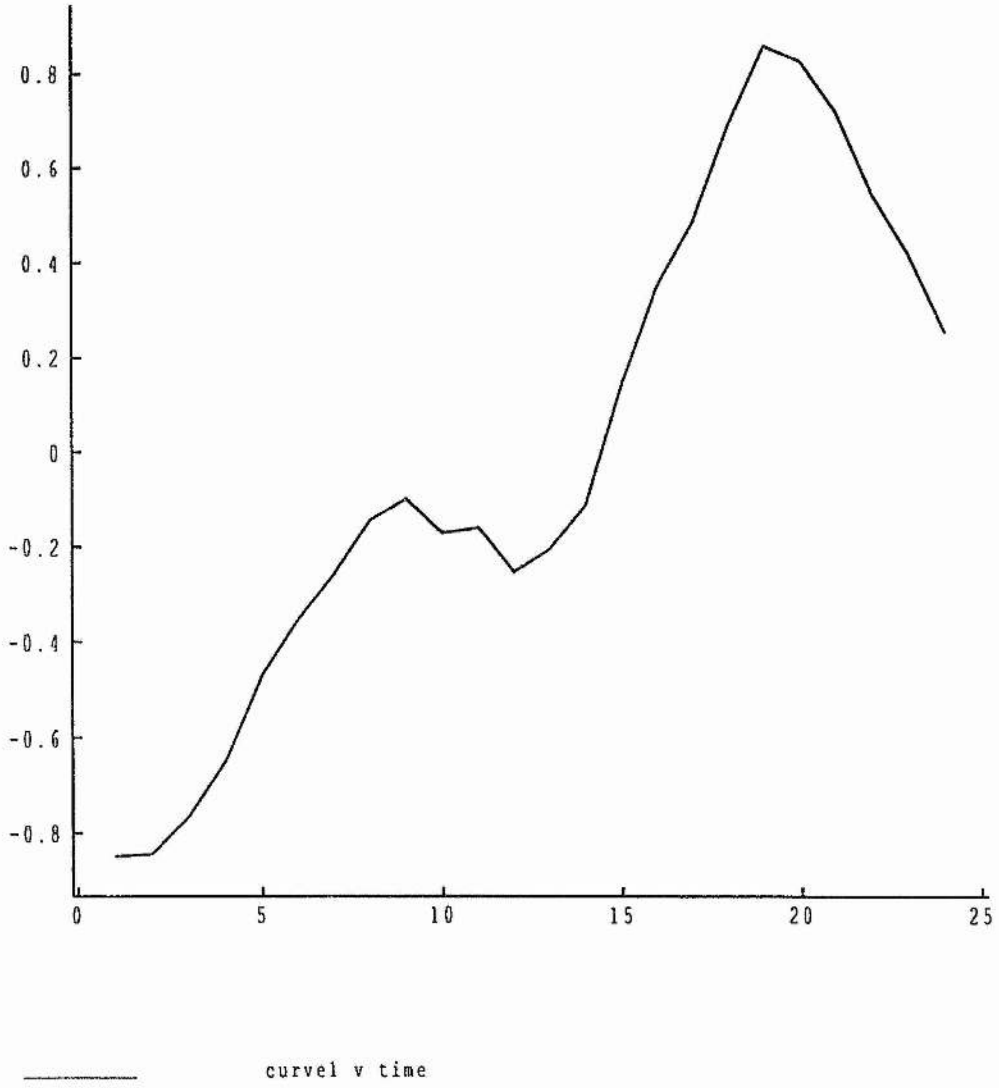


Fig 6.25 (Part 3): Ozone Curve: Principal Point PP2

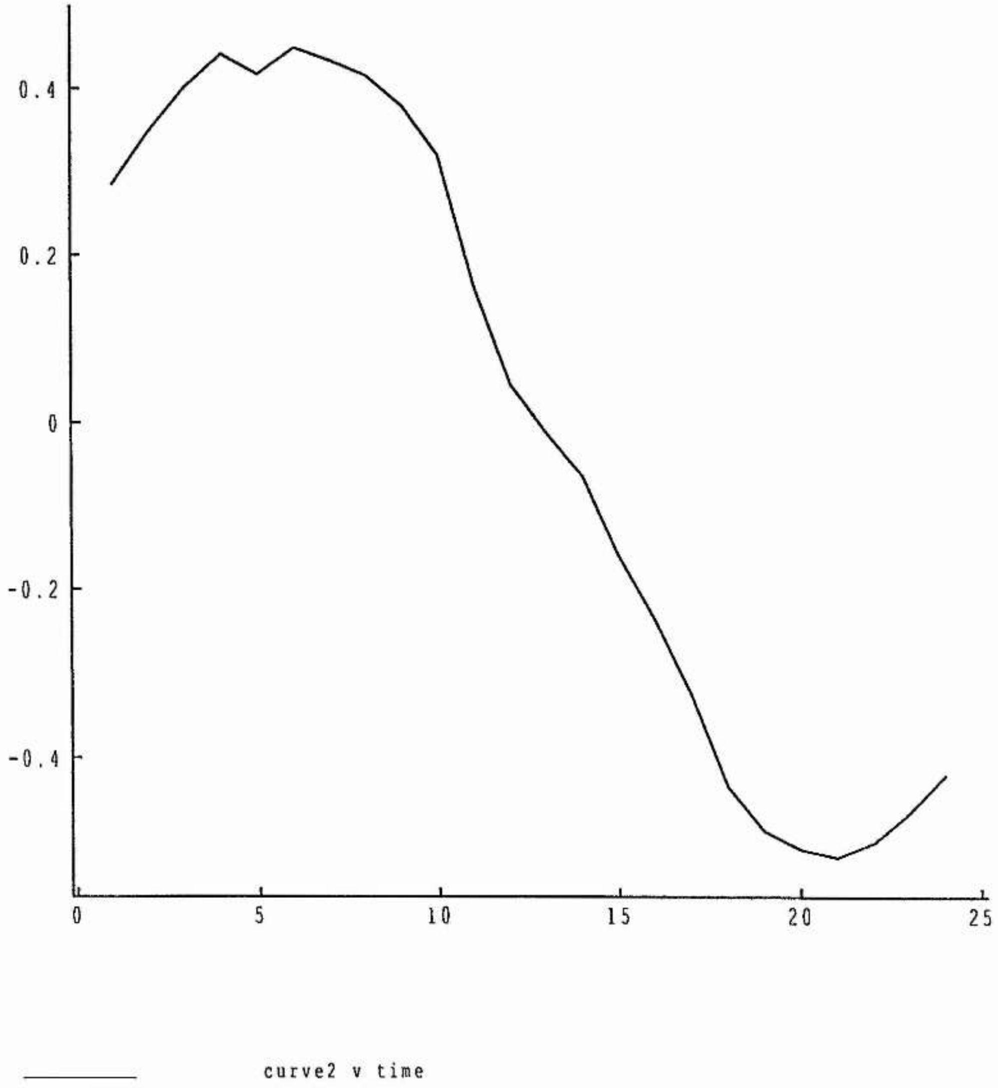
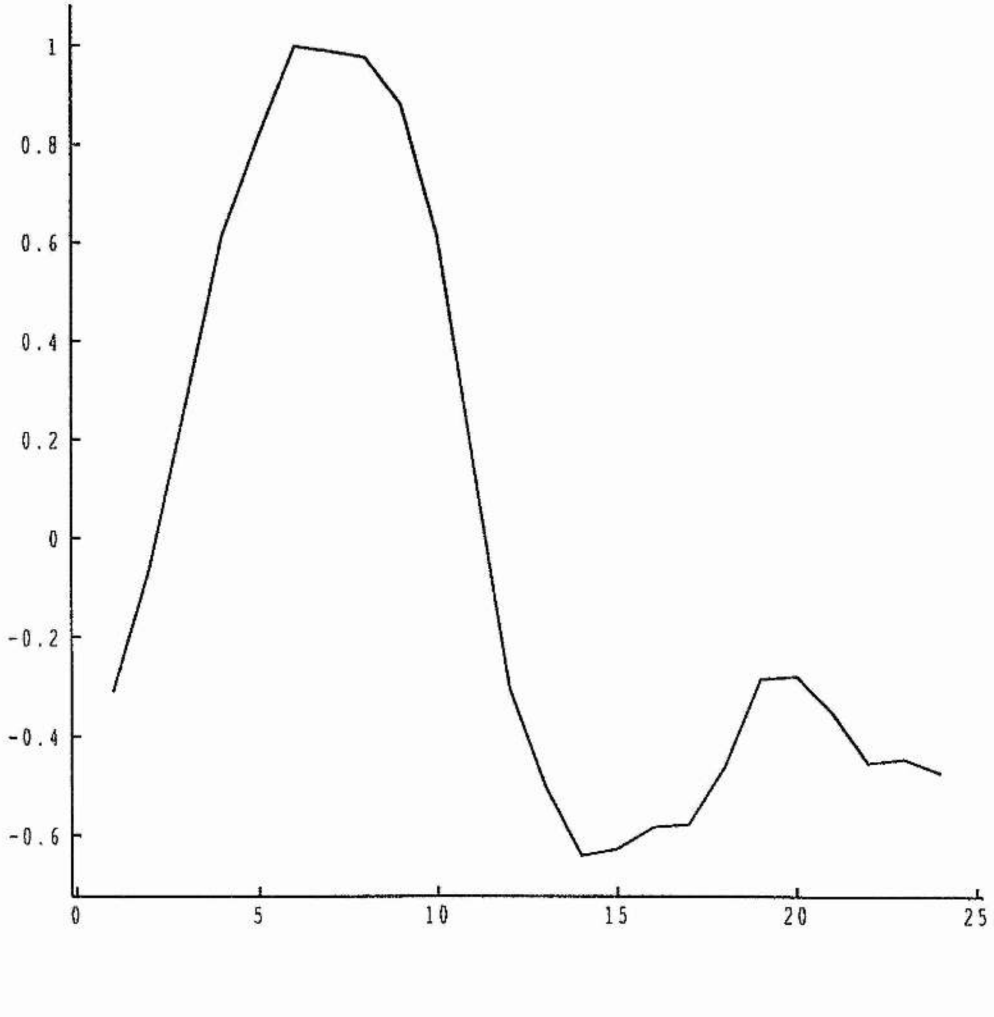


Fig 6.25 (Part 4): Ozone Curve: Whole Day WD1



Chapter 7: Conclusion

In chapter 2, the inverse Gaussian distribution was considered as a possible alternative to the lognormal as a model for the frequency distribution of observed SO_2 concentrations. It was suggested that as the inverse Gaussian distribution has a heavier upper tail than the lognormal, it might have been possible to address the problem of underestimation of the frequency of potentially damaging high concentrations of SO_2 by its use. The work carried out indicated that the fit of the inverse Gaussian was even less acceptable than the lognormal. Given a better result at this stage, it would have been possible to improve the fit by using alternative fitting methods such as a minimisation of the Anderson-Darling test statistic considered, but with the results obtained this was not worthwhile.

Chapter 3 made use of the fact that data were available from two ozone analysers recording side by side to investigate any differences between the observations obtained from both machines. The data analysis indicated that while the actual value of the differences was small (generally only 1-2 ppb) there were a number of occasions in every sequence where only one of the two machines had recorded a temporary jump in ozone levels, by as much as 40ppb. The fact that this jump only occurred on one of the two machines, and occurred on each with roughly equal frequency for each, suggests that these jumps are a feature of the monitoring equipment. These jumps could have serious implications for analyses of extreme values conducted on these data, as the jumps would in many cases be considered to be the daily or hourly maxima. It would therefore be of great interest to obtain more data sets where two machines are available and investigate the occurrences of these jumps further. Removing the jumps in the data, three Box-Jenkins ARIMA type models were considered as possible models for the difference in readings between the two machines. One of these, the IMA(1,1) model, was shown to have a sound theoretical basis for selection, and fitted the data reasonably well. This approach could be extended by the use of a similar model, but with a non-Gaussian random shock distribution, as the residuals obtained after model fitting were shown to be from a distribution with slightly heavier tails than the normal. The t-distribution is an obvious choice, and this would be the logical next step in any extension of this work.

Chapter 4 began the largest topic of this thesis, the analysis of the behaviour of ozone levels over a day. This first step was to consider the average diurnal cycle of ozone for each month of a year, and investigate the variation in shape and level of these curves. This chapter also introduces the concept of distorting the time axis for each day of the year such that sunrise, sunset, midday and midnight occurred at the same (transformed) time each day. This approach was shown to make comparison of the behaviour of ozone in a day over the year simpler, as any effects due to the position of the sun in the sky (notably solar radiation and temperature) are not masked by changes in sunrise and sunset times over the year. The monthly average diurnal curves calculated after this time distortion was carried out were shown to have similar properties, with the daytime and night time sections being quite distinct in each case. The reduced parameter version of the polynomial model was shown to provide a reasonable description of these curves from month to month.

The diurnal behaviour of ozone was further investigated in chapter 5. The difference in average diurnal cycle from month to month over a year was hypothesised to be due to the existence of several typical types of ozone day, representing different weather conditions, whose relative frequencies change over a year. The work carried out in chapter 5 sought to identify these types, by the use of a classification methodology treating each ozone day as an object for classification. Due to the work carried out in chapter 4, it was decided that the use of the distorted time axis was desirable when considering days from different times of the year together. It was also decided after the work in chapter 4 to consider the whole day, daytime and night-time curves from each day separately. The classification methodology was then used and identified two types of whole day curve, four types of daytime curve and five types of night time curve. The behaviour of ozone over the daytime, night time or whole day curve was shown to be different in level and/or spread according to which basic type the curve had been described as. The distribution of weather types during the day, night or whole day was also shown to be significantly different over types. It was also shown to be possible to explain much of the variability observed in the monthly average curves considered in chapter 4 by using the frequency of types of whole day observed in each month. These facts indicate that the different types of curve discovered by the classification method can be related to differences in physical processes and that they can reasonably be considered to have physical meaning. As further validation of this approach, data from another site were considered using the same methodology, and the resulting whole day shapes were found to be very similar. This provides further evidence that the types discovered represent differing physical processes.

There are many possibilities for further work using this classification methodology. First, data from many other sites could be analysed. If similar types are discovered in the data from these, then further investigation of the behaviour of ozone within each of these could prove useful for understanding the different processes governing ozone production and depletion. Each type could possibly be related to a different set of physical processes operating. Modelling ozone behaviour within the same type could be made simpler, as each curve would be generated by similar processes. The relative frequencies of the types over time or different sites could also provide valuable insights into the changes in physical processes governing ozone levels.

The classification methodology introduced could also prove useful when analysing data from other areas. The primary case would be when data are in the form of objects consisting of ordered observations. The 'area between two curves' dissimilarity measure has been shown in chapter 6 to have desirable theoretical properties, and takes account of any such ordering, unlike most measures of dissimilarity commonly used. The importance of considering the ordering was illustrated in section 6.1, where examples are given where ignoring the time ordering by using a standard dissimilarity measure leads to very different conclusions.

References

Chhikara, R.S. & Folks, J.L (1989), *The Inverse Gaussian Distribution: Theory, Methodology and Applications.* New York, Marcel Dekker

Filliben, J. (1975), The Probability Plot Correlation Coefficient Test for Normality, *Technometrics*, **17**, 111-117.

Flury, B. (1990), Principal points. *Biometrika*, **77**, 33-41.

Flury, B. (1993), Estimation of principal points. *Applied Statistics*, **42**, 139-151.

Fowler, D. & Cape, J.N. (1982), Air Pollutants In Agriculture and Horticulture. In *Effects of Gaseous Air Pollution in Agriculture and Horticulture* (eds M.H.Unsworth and D.P.Ormrod), pp.3-26, London: Butterworth.

Gordon, A. D. (1987), *Classification* . London, Chapman & Hall

Hartigan, J. A. (1975), *Clustering Algorithms*. New York, Wiley - Interscience.

Lance, G.N. & Williams, W.T. (1967), A general theory of classificatory sorting strategies. I. Hierarchical systems. *Computer J.* , **9**, 373-380.

Larsen, K.A. (1973), An Air Quality Data Analysis System for Interrelating Effects, Standards and Needed Source Reductions, Part 1. *Journal of the Air Pollution Control Association*, **23**, 933 - 940.

Larsen, K.A. (1974), An Air Quality Data Analysis System for Interrelating Effects, Standards and Needed Source Reductions, Part 2. *Journal of the Air Pollution Control Association*, **24**, 551 - 558

Larsen, K.A. (1976), An Air Quality Data Analysis System for Interrelating Effects, Standards and Needed Source Reductions, Part 3. *Journal of the Air Pollution Control Association*, **26**, 325 - 333

Larsen, K.A. (1977), An Air Quality Data Analysis System for Interrelating Effects, Standards and Needed Source Reductions, Part 4. *Journal of the Air Pollution Control Association*, **27**, 454 - 459.

Rawlings, J.O. & Cure, W.W. (1985); The Weibull Function as a Dose-Response Model to Describe Ozone Effects on Crop Yields. *Crop Science*, **25**, 807 - 814.

Rawlings, J.O., Lesser, V.M. & Dassel, K.A. (1988), Statistical Approaches to Assessing Crop Losses. In *Assessment of Crop Loss from Air Pollutants* (eds W.W. Heck, O.C. Taylor & D.T. Tingey), pp. 389-416. London, Elsevier Applied Science.

Sinclair, C.D., Spurr, B.D. & Ahmad, M.I. (1990), Modified Anderson Darling Test. *Comm. in Statistics, Theory & Methods*, **19** , 3677-3686.

Smith, R.I., Fowler, D. & Cape, J.N. (1989), The Statistics of Phytotoxic Air Pollutants. *J.R.Statist. Soc. A* ., **152**, 183-198.

United Kingdom Photochemical Oxidants Review Group (1987), *Ozone in the United Kingdom.*, Interim Report. London, Department of the Environment.

United Kingdom Photochemical Oxidants Review Group (1990), *Oxides of Nitrogen in the United Kingdom.*, Second Report. London, Department of the Environment.

United Kingdom Photochemical Oxidants Review Group (1993), *Ozone in the United Kingdom 1993.*, Third Report . London, Department of the Environment.

Wei, W.W.S (1990), ,*Time Series Analysis, Univariate and Multivariate Methods* . Addison-Wesley. (Model Selection chaps 6&7 - Aggregation chap 16)

Appendix 1: Program to fit quintic model

```
*******************************************
* Program to fit 40 parameter 2-quintic model to data *
* (c) 16/4/93 Paul Hutchison                          *
*******************************************


C
C Definitions
C
      implicit real*8 (a-h,o-z)
       dimension y(96,12),f(8,96),alp(16),X(96,2),XT(2,96),XTX(2,2)
       DIMENSION XVX(2,96),BE(2,12),AK(16),AM(16,16),Q(16),G(16),AMK(16)
       dimension me(16),mf(16),obe(2,12),oalp(16),vxtx(2,2)
      character*7 na(12)
       open(1,file='na.dat',status='old')
       read(1,*) na
       close(1)
C
C Input of initial values for alpha 1-8 and beta 1-8
C
       TYPE '(1X,A,$)','Please input the initial values of ALP:'
       read(*,*) alp
       write(*,*) 'Thank you - please wait'
       do 5 i=1,16
       oalp(i)=alp(i)
5      continue


C
```

```
C Input of data
C
      iteration=0


      do 20 i=1,12
      open(1,file=na(i),status='old')
      do 10 j=1,96
      read(1,*) a
      y(j,i)=a
10     continue
      close(1)
20    continue

C
C Calculation of the Fkj polynomials
C
      do 30 i=1,96
      ti=-1+(i-1)*2/96.d0
      ti2=ti*ti
      ti3=ti2*ti
      ti4=ti3*ti
      ti5=ti4*ti
      tip2=0
      if(ti.gt.0.) tip2=ti2
      tip3=tip2*ti
      tip4=tip3*ti
      tip5=tip4*ti
      f(1,i)=1
      f(2,i)=ti3-ti
      f(3,i)=ti4-2*ti2
      f(4,i)=ti5-ti
      f(5,i)=tip2-.5*ti-.5*ti2
      f(6,i)=tip3-.5*ti-.75*ti2
      f(7,i)=tip4-.5*ti-ti2
      f(8,i)=tip5-.5*ti-1.25*ti2
30    continue

C
C Main loop - calculation of new estimates
C

34    do 35 i=1,2
      do 35 j=1,12
      obe(i,j)=be(i,j)
35    continue

      do 50 i=1,96
      do 50 j=1,2
      nz=1+(j-1)*8
      X(I,j)=0
      do 40 k=1,8
       X(I,J)=X(I,J)+ALP(NZ)*F(k,i)
      NZ=NZ+1
40    continue
50     XT(j,i)=x(i,j)

      do 60 i=1,2
      do 60 j=1,2
      xtx(I,J)=0
      do 60 k=1,96
      xtx(i,j)=XT(i,k)*X(k,j)+XTX(I,J)
60    continue

      delta=xtx(1,1)*xtx(2,2)-xtx(1,2)*xtx(2,1)
      vxtx(1,1)=xtx(2,2)/delta
       vxtx(1,2)=-xtx(2,1)/delta
       vxtx(2,1)=-xtx(1,2)/delta
       vxtx(2,2)=xtx(1,1)/delta

       do 70 i=1,2
       do 70 j=1,96
       xvx(i,j)=0
       do 70 k=1,2
       xvx(i,j)=xvx(I,J)+vxtx(i,k)*xt(k,j)
70    continue

       do 80 i=1,2
       do 80 j=1,12
       be(i,j)=0
       do 80 k=1,96
       be(i,j)=be(I,J)+xvx(I,k)*y(K,j)
80    continue

       do 85 i=1,16
       oalp(i)=alp(i)
85    continue
```

```
      do 100 i=1,2
      do 100 j=1,8
      su=0
      do 90 k=1,96
      do 90 l=1,12
      su=su+y(K,l)*BE(I,L)*F(J,K)
90    continue
      nd=j+(i-1)*8
      AK(nd)=su
100   continue

      do 120 i=1,2
      do 120 j=1,2
      do 120 k=1,8
      do 120 L=1,8
      su=0
      do 110 m=1,12
      do 110 n=1,96
      su=su+BE(I,M)*BE(j,M)*F(K,N)*F(L,N)
110   continue
      NR=K+(i-1)*8
      NS=L+(j-1)*8
      AM(NR,NS)=su
120   continue

      call ivsnc(AM,Q,G,ME,MF,16,1.D-20)

      do 130 i=1,16
      AMK(i)=0
      do 130 j=1,16
      AMK(i)=amk(i)+AM(i,j)*AK(J)
130   continue

      do 140 i=1,16
      alp(i)=amk(i)
140   continue

      iteration=iteration+1
      ss=0
      do 146 i=1,12
      do 146 j=1,96
      sq=y(j,i)
      do 145 k=1,2
      do 145 l=1,8
      sq=sq-be(k,i)*alp(l+(k-1)*8)*f(l,j)
145   continue
      ss=ss+sq*sq
146   continue
      if(iteration.eq.1) go to 34
C
C New estimates made - Check for convergence
C
      lag=0
      do 150 i=1,2
      do 150 j=1,12
      test=be(i,j)-obe(i,j)
      test=test/be(i,j)
150   if(test.gt.0.0001) lag=1
```

```
      do 160 i=1,16
      test=alp(i)-oalp(i)
      test=test/alp(i)
160   if(test.gt.0.0001) lag=1

      write(*,*) 'Iteration number:',iteration,'ss:',ss

      if(lag.eq.1) go to 34
C
C Output results
C
      write(*,*)
      write(*,*) 'Convergence Criterion Met after ',iteration,' iterations'
      write(*,*)
      write(*,*) 'Values of ALP (alpha1-8,beta1-8)'
      write(*,*)
      do 165 i=1,8
      write(*,*) i,'  ',alp(i),'      ',alp(i+8)
165   continue
      write(*,*)
      write(*,*) 'Values of BE (phi1-12,xi1-12)'
      write(*,*)
      do 166 j=1,12
      write(*,*) j,'  ',be(1,j),'      ',be(2,j)
166   continue

      stop
      end
*
```

```
* subroutine to invert a matrix
*
      SUBROUTINE IVSNC(A,B,C,ME,MF,N,EP)
      implicit real*8 (a-h,o-z)
      DIMENSION A(N,N),B(N),C(N),ME(N),MF(N)
      DO 10 K=1,N
      Y=0.
      DO 20 I=K,N
      DO 20 J=K,N
      IF(dABS(A(I,J)).LE.dABS(Y)) GO TO 20
      Y=A(I,J)
      I2=I
      J2=J
20    CONTINUE
      IF (dABS(Y).LT.EP) GO TO 32
      IF (I2.EQ.K) GO TO 33
      DO 11 J=1,N
      W=A(I2,J)
      A(I2,J)=A(K,J)
11    A(K,J)=W
33    IF (J2.EQ.K) GO TO 44
      DO 22 I=1,N
      W=A(I,J2)
      A(I,J2)=A(I,K)
22    A(I,K)=W
44    ME(K)=I2
      MF(K)=J2
      DO 50 J=1,N
      IF (J-K) 2,3,2
```

```
3     B(J)=1./Y
      C(J)=1.
      GO TO 4
2     B(J)=-A(K,J)/Y
      C(J)=A(J,K)
4     A(K,J)=0.
      A(J,K)=0.
50    CONTINUE
      DO 40 I=1,N
      DO 40 J=1,N
40    A(I,J)=A(I,J)+C(I)*B(J)
10    CONTINUE
      DO 60 L=1,N
      K=N-L+1
      K1=ME(K)
      K2=MF(K)
      IF(K1.EQ.K) GO TO 70
      DO 55 I=1,N
      W=A(I,K1)
      A(I,K1)=A(I,K)
55    A(I,K)=W
70    IF(K2.EQ.K) GO TO 60
      DO 66 J=1,N
      W=A(K2,J)
      A(K2,J)=A(K,J)
66    A(K,J)=W
60    CONTINUE
      RETURN
32    EP=-EP
```

```
RETURN

END
```

Appendix 2: Program to carry out classification

```
program ClassifyDays
*
* Input of parameters and data
*

      double precision e,f,g,h,minx,dij,w,o,d,d1,str,dm,zx,bx
      integer c,n,m,i,j,k,mini,minj,l,s,lx,a,b
      character*8 ans,fname
C
C Dimension arrays
C

      dimension w(100),o(1000,1001),d(1000,100),d1(100),dm(1000,100)
      dimension bx(1000)

C
C Version message
C

      write(*,*)
      write(*,*) '******************************************'
      write(*,*) '*                                *'
      write(*,*) '* Ozone days classification - Ver 8      *'
      write(*,*) '* Mean Curve Version                   *'
      write(*,*) '* (c) 24th Aug 1993 P.Hutchison          *'
      write(*,*) '*                                *'
      write(*,*) '****************************************'


C
C N - total number of objects
C

      write(*,*)
      write(*,*) 'Total number of days? (Max 1000)'
      read(*,*) n

C
C C - Max number of clusters required
C

      write(*,*)
      write(*,*) 'Max number of clusters required?'
      read(*,*) c

C
C A - Min number of clusters required
C

      write(*,*)
      write(*,*) 'Min number of clusters required?'
      read (*,*) a
C
C M - number of observations per object
C
```

```
      write(*,*)
      write(*,*) 'Number of observations per day? (Max 100)'
      read(*,*) m


C
C W(m-1) - Weight funtion (W(i) is weight for area between obs i & i+1)
C

      write(*,*)
      write(*,*) 'Do you wish to enter a weight function? (y/n)'
      read(*,500) ans
      if (ans.eq.('y')) then
         write(*,*)
         write(*,*) 'Weight function :',m-1,'weights required'
         write(*,*) 'Enter filename for read (max 8 chars)'
         read(*,500) fname
         open(1,file=fname,status='old')
         do 10 i=1,m-1
            read(1,*) w(i)
10       continue
         close(1)
      else
         do 15 i=1,m-1
            w(i)=1
15       continue
      endif
```

```
C
C Standardisation used in calculating dij
C 1-mean 2-sd 3-mean+sd 4-none
C

      write(*,*)
      write(*,*) 'Standardisation to be used before clustering'
      write(*,*)
      write(*,*) '1 - mean'
      write(*,*) '2 - sd'
      write(*,*) '3 - mean & sd'
      write(*,*) '4 - none'
      write(*,*)
      write(*,*) 'Enter choice (1-4)'
      read(*,*) s

C
C D(n,m) - data
C
      write(*,*)
      write(*,*) 'Data input'
      write(*,*) 'Enter filename for data read (max 8 chars)'
      read(*,500) fname
      open(2,file=fname,status='old')
      do 20 i=1,n
      do 20 j=1,m
         read(2,*) d(i,j)
         dm(i,j)=d(i,j)
 20   continue
```

```
      close(2)

C
C Output filename
C

      write(*,*)
      write(*,*) 'Enter filename for output (max 8 chars)'
      read(*,500) fname
      open(3,file=fname,status='new')

C
C Standardise data
C
      if (s.ne.4) then
         do 26 i=1,n
         do 25 j=1,m
         d1(j)=d(i,j)
25       continue
         call standardise(d1,s,m)
         do 26 j=1,m
         d(i,j)=d1(j)
         dm(i,j)=d(i,j)
26       continue
      endif

C
C Initialise clusters
C
```

```
      do 30 i=1,n
          o(i,1)=1
          o(i,2)=i
30    continue

*
* Calculate stress before any clustering occurs
*

C
C Calculate Overall mean curve
C

      do 40 i=1,m
      d1(i)=0
40    continue
      do 50 i=1,n
      do 50 j=1,m
      d1(j)=d1(j)+d(i,j)/n
50    continue

C
C Calculate Mean Dij
C

      str=0
      do 70 j=1,m-1
      do 70 i=1,n
      e=d1(j)
      f=d1(j+1)
      g=d(i,j)
      h=d(i,j+1)
      call area(e,f,g,h,zx)
      str=str+zx*w(j)
70    continue
      str=str/n
      write(3,*)
      write(3,*) 'Stress before clustering'
      write(3,*) str
      write(3,*)

*
* Do the following procedures until the correct number of clusters is reached
*
      do 140 l=a,n-1
*
*  Decide which 2 clusters should be combined
*

      minx=-99
      mini=0
      minj=0
      do 110 i=2,n
      do 110 j=1,i-1
C
C If clusters i and j are non-empty then calculate dij
C w(k) is the weight function
```

```
C
      if (o(i,1).ne.0.and.o(j,1).ne.0) then
         dij=0
         do 100 k=1,m-1
         e=d(i,k)
         f=d(i,1+k)
         g=d(j,k)
         h=d(j,1+k)
         call area(e,f,g,h,zx)
         dij=dij+zx*w(k)
100      continue

C
C check to see if this is the minimum dij
C
         if (minx.eq.-99.or.dij.lt.minx) then
            minx=dij
            mini=i
            minj=j
         endif
      endif
110   continue

C
C Iteration message
C

      write(3,*) '*******************************************************'
      write(3,*)
      write(3,*) n-l+a,' clusters - combine ',mini,minj,' dij = ',minx
      write(3,*)
      write(*,*) n-l+a

*
* Combine the 2 clusters chosen
*

C
C Add element list of cluster j to list for i
C
      ni=o(mini,1)
      nj=o(minj,1)
      do 120 i=1,nj
      o(mini,ni+1+i)=o(minj,i+1)
120   continue
      o(mini,1)=ni+nj
C
C Update mean of cluster i
C
      do 130 i=1,m
      d(mini,i)=(d(mini,i)*ni+d(minj,i)*nj)/(ni+nj)
130   continue

C
C Set cluster j to empty
C
      o(minj,1)=0
```

```
*
* End of clustering procedure
*

      if ((n-l+a).le.c+1) then
*
* Output results
*

      lx=1
      do 150 i=1,n
      if (o(i,1).ne.0) then
C
C Cluster number
C
         write(3,*) 'cluster ',lx
         lx=lx+1
C
C Number of objects in cluster
C
         write (3,*) 'size ',o(i,1)
C
C     list of objects in cluster
C
         write(3,*) 'object list'
         do 160 j=1,o(i,1)
            bx(j)=o(i,j+1)
160      continue
```

```
          b=idint(o(i,1))
          call sort(b,bx)
          do 165 j=1,o(i,1)
             write(3,*) bx(j)
165       continue


C
C    cluster mean object
C
          write(3,*) 'cluster mean'
          do 170 j=1,m
             write(3,*) d(i,j)
170       continue
C
C Cluster 'stress'
C
       write(3,*) 'cluster stress'
       str=0
       do 180 k=1,o(i,1)
       do 180 j=1,m-1
          zx=o(i,k+1)
          e=d(i,j)
          f=d(i,1+j)
          g=dm(zx,j)
          h=dm(zx,1+j)
          call area(e,f,g,h,zx)
          str=str+zx*w(k)
180    continue
       str=str/o(i,1)
```

```
          do 210 i=1,m
              s1=s1+d1(i)*d1(i)
210       continue
          s1=1.d0/(m-1)*(s1-m1*m*m1)
          s1=sqrt(s1)
      endif
      if (s.eq.2) then
          m1=0
      endif
      if (s1.eq.0) then
          s1=1
      endif
      do 220 i=1,m
          d1(i)=(d1(i)-m1)/s1
220   continue
      return
      end

*
* Procedure to calculate area of polygon formed from a,b,c,d
* (corrupts a,b,c,d)
*

      subroutine area(a,b,c,d,x)

      double precision a,b,c,d,x
      x=min(a,b,c,d)
      a=a+x
      b=b+x
```

```
       if (o(i,1).eq.1) then
       str=0
       endif
       write(3,*) str
       endif
 150   continue
       endif
 140   continue
       close(3)
 500   format(a)
       end



*
* Procedure to standardise a day held in d1 by me,sd,me+sd s=1,2,3
*


       subroutine standardise(d1,s,m)


       double precision m1,s1,d1(100)
       integer s,m
       m1=0
       s1=1
       do 200 i=1,m
       m1=m1+d1(i)
 200   continue
       m1=m1/m
       if (s.ne.1) then
          s1=0
```

```
      c=c+x
      d=d+x
      if (sign(1,a-c).eq.sign(1,b-d)) then
         x=dabs(0.5*(a+b)-0.5*(c+d))
      elseif ((d-c+a-b).ne.0) then
         x=dabs(0.5*((c-a)*(c-a)+(d-b)*(d-b))/(d-c+a-b))
      else
         x=0
      endif
      return
      end
*
* Subroutine to sort the n elements of vector d
*

      subroutine sort(n,d)

      double precision d(1000),temp
      integer n,i,j
      do 300 j=1,(n-1)
      do 300 i=1,(n-j)
      if (d(i).gt.d(i+1)) then
         temp=d(i)
         d(i)=d(i+1)
         d(i+1)=temp
      endif
300   continue
      return
      end
```