Prediction of Solar and Geomagnetic Activity Using Artificial Neural Networks

by

Keith P. Macpherson B.Sc.

Thesis
submitted to the
University of Glasgow
for the degree of
Ph.D.

Astronomy and Astrophysics Group
Department of Physics and Astronomy,
University of Glasgow,
Glasgow G12 8QQ

December 1994

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

With thanks to my parents.

For my sister, a *real* doctor and an inspiration.

Acknowledgements

During my time as a postgraduate research student I have been fortunate not to have been locked up in a room and told only to come out in three years with some results (much as some people might like the former idea). Instead I have had the opportunity to travel to conferences, workshops and research meetings to discuss work and ideas. These trips have enhanced not only my research but also my life, and for this I thank my supervisor Professor John Brown. His encouragement to his students to see their research as part of the wider scientific community is to be commended and the renewed enthusiasm which a student gets out of research trips is not to be underestimated.

My research was carried out while in receipt of a Science and Engineering Research Council studentship. SERC also provided financial assistance to allow me to participate in international workshops and conferences. I would also like to thank them for granting an extension to my funding in the light of personal circumstances. Additional funding from the European Space Operations Centre is also greatly appreciated.

During the seven and a half years that I have been studying in the Astronomy and Astrophysics Group of the Department of Physics and Astronomy, I have come to appreciate the friendly and relaxed atmosphere which usually exists (removals excepted) in this Group. I would like to take this opportunity to thank all those people who have contributed at some point to this atmosphere, be they permanent features or transient phenomena.

There is also, however, a list of people whom it would be remiss of me not to mention personally. At the very top of this list is Andy Newsam, without whom none of this would have been possible. When Andy arrived here after my first year and I had never written a computer program single-handed, little would I believe that I could complete such an intensely computational thesis. The fact that I did is down to Andy's genius and a fair amount of his patience as well. So if you decide you want my LaTeX style file for your thesis, Andy, then I guess I cannot begrudge you that. On the computing aspect of affairs I also have to thank Norman for deciphering the mysteries of TeX and LaTeX (in particular for aforementioned style file) as well as for keeping all the suns ticking over. Latterly this responsibilty passed to Shashi Kanbur (Captain Starlink), who so far has succeeded in answering my calls for ever-increasing amounts of diskspace. In the last 18 months, much of this research has been greatly enhanced through discussions with Andrew Conway. A shared passion

for neural networks during work hours (maybe) and certainly for malt whisky at other times gave rise to memorable times in Sweden and Germany. The thoughts and comments from Guy Janin of ESOC, along with his continual interest in the project, has also been immensely beneficial to the work. It is also impossible to imagine what life would be like in this department without the presence of Daphne to sort out all things non-scientific.

I would also like to make a special mention for all the people who played their part in making conference visits such enjoyable experiences instead of daunting prospects. In particular are David and Colin for COSPAR and the IAF, Andrew and Peter in Lund and Lyndsay, Giota and Peter (again) for SAAS-FEE.

In recent months it has been hard sometimes to remember life outside of a thesis. I would like to take this time to show my appreciation to the many people whose friendships I sometimes take for granted, but whose existence maintains the balance in my life. In no particular order (except alphabetical), I mention Allison, Andy, Iain, Iain, Martin, Neil, Stuart, Tony and Wendy although there are several others who could be added to this list. It is no surprise perhaps that many of those listed above share my appreciation of good malt whisky.

Lastly, I would like to thank three people who have played a most significant role in my life to date. To John Taylor, I greatly appreciate the provision (above and beyond the call of duty) of tea, toast and a sofa to sleep on, as well as introducing me to the 'benefits' of popmobility and tune-ups. Although several hundred miles away, my friend Marion has long been my closest companion with respect to thesis blues and life in general. To complete this threesome, I thank Mark for fulfilling at various times the role of playing partner, opponent, team captain, brother and a true and loyal friend.

Finally, to my family, I say that I cannot adequately express in words my thanks for all your support. The best I feel I can do is to live my life to the full, since this was your gift to me and I wish to show that I truly appreciate it.

Summary

The work of this thesis is concerned with investigating the application of artificial neural network techniques to the problem of predicting the pattern of behaviour of the solar activity cycle. This investigation measures the success with which these computational methods can predict various solar activity indices on different timescales and in each case compares the level of success with that obtained using a comparison benchmark prediction model.

Taken individually, the two subjects involved in this thesis, solar activity prediction and artificial neural network methods, are enormous areas of active research. The overlap, however, has a very limited history and this thesis aims to redress some of these limitations. The most important aspect of my research has been to produce the most accurate predictions of future solar activity possible. In doing this, it is essential to compare the prediction accuracy achieved in my analysis with that currently provided by other methods. In carrying out this research though, I have investigated in detail the effect of varying the parameters which define a neural network on the prediction accuracy achievable. In this way I hope to go some way towards bridging the gap between the two different subject areas.

The general outline of the thesis is as follows. The opening two chapters describe in turn the relevant history of solar activity prediction and basic neural network theory. Included in Chapter 1 is a historical perspective of the observations of solar activity phenomena concentrating on early measurements of sunspots. From these measurements the approximate 11-year solar cycle was discovered. The implications of high levels of solar activity for satellite drag, solar particle events and geomagnetic storms are detailed to demonstrate some of the observed interactions between the Sun, the interplanetary medium and the terrestrial environment. A brief review of the types of prediction models which are currently in use for estimating the future behaviour of the Sun is also carried out. Neural network computational techniques were born over fifty years ago but have experienced mixed fortunes in terms of popularity during these years. Since the mid-1980's interest has increased markedly in these methods, in the main due to the emergence of the 'back-propagation of errors' learning algorithm. This algorithm is described in Chapter 2 along with the historical perspective which gave rise to its formulation.

Since this research is essentially numerical in nature, a link has to be provided between the theory

and the practical implementation. This is done in Chapter 3 where the methodology required to analyse the time series solar activity data using the back-propagation algorithm is described. In particular, the software developed in the course of this research is described, showing how it relates to the back-propagation algorithm. The limited amount of previous work in this subject area is also described along with consideration of potential problems in neural network learning which can occur in a practical implementation of the theory.

Chapters 4, 5 and 6 collectively discuss the results of the comprehensive investigations which have been carried out for this thesis. Chapter 4 is devoted solely to analysis of the smoothed monthly sunspot number. This is because this index has been most widely used as an indicator of solar activity levels and has been observed consistently for the longest time. A full description of the effect of changing the various parameters in a neural network is also given. The next chapter summarises the results obtained when a similar analysis is applied to unsmoothed monthly sunspot numbers, monthly and yearly solar flux data and also geomagnetic data in the form of the antipodal *aa* index. Thereafter, Chapter 5 discusses some alternative ways of presenting data as input to the networks, for example, a combination of neural network training with the McNish and Lincoln method is tested. Different styles of obtaining a prediction for n sample points into the future are also established. Finally, the last half of this chapter concentrates on the specific problem of predicting only the level of activity at the next maximum. This builds on some ideas previously suggested by Koons and Gorney (1990) and compares them with the procedures studied in the earlier parts of this thesis.

Chapter 7 offers a brief overview of the results and conclusions of the previous chapters and completes this with a look at some of the prospective areas into which this research may be extended. The original work of this thesis is contained in Chapters 3 through 6 which detail the various investigations carried out into the application of artificial neural network methods to solar activity prediction. The initial results of these studies have been presented at various international conferences or workshops and appeared in subsequent proceedings. The more detailed and complete results are in preparation for submission for publication as full papers.

Contents

Chapter 1

The Nature of Solar Activity

1.1 Introduction

Although in many astronomy circles, the field of solar physics is not considered part of astronomy and astrophysics, the study of our star, the Sun, is, I believe, of extreme importance for two reasons. The first is due to our reliance on it for life-preserving energy and the effect that any fluctuations in the amount of energy radiated have on the terrestrial environment. The second is that by virtue of our closeness to the Sun, we are in a privileged position to study its complex physical behaviour. Indeed despite close observation from ground- and space-based instrumentation, there is a great deal which is still to be learnt and understood about many of the Sun's physical mechanisms. The work of this thesis is concerned with methods for analysing the variations in solar activity which are observed. In this way, it is hoped that the subsequent effects on the terrestrial environment may be predicted to a greater level of accuracy than currently exists. Due to the sheer complexity of the Sun-Interplanetary Medium-Earth interactions, and a lack of a mechanism to explain the observed behaviour, the methods under study here resort to identification of past patterns of behaviour and correlations between events.

In this chapter a history and description of the fluctuations of activity observed on the Sun will be given. Consideration will be given to the form of the measurements which best describe the pattern of behaviour of solar activity. These solar indices, along with parallel measurements made of the variations in the Earth's magnetic field, have been recorded for many years and are available on a daily, monthly and yearly basis. This chapter will also be concerned with describing the known effects of variations in solar activity on the terrestrial environment and the importance attached to an improvement of our level of understanding in this area. Finally, a description of current, established techniques for analysing and predicting solar and geomagnetic indices will be provided. One of these techniques in particular will be used to provide the standard against which the methods outlined in this thesis will be compared.

1.2 Nature of Solar Activity

The levels of activity exhibited by the Sun can be appreciated in two main ways. One is through consideration of the various transient phenomena which are observed in the solar atmosphere, either by the naked eye or through instrumentation. These transient features consist of sunspots, solar flares, prominences and general active regions which are superimposed on the Quiet Sun background. High levels of occurrence of these structures and events represent an active Sun. All of these phenomena owe their existence to the presence of magnetic fields in the solar atmosphere. When viewed in white light, most of these phenomena are not visible. When the photosphere and overlying atmosphere is specifically studied, in for example $H\alpha$ wavelength, the details are observed. For the purposes of this thesis, most of our attention will be concerned with describing the behaviour of sunspots since, as will become clear from the text, it is the existing recorded variation in the number of sunspots which currently gives the longest standing detailed record of solar activity.

Alternatively, the level of activity of the Sun can be measured through spectrophotometry of the whole solar energy output. Most of the Sun's energy is in the form of low energy photons in the ultraviolet through infra-red regions of the spectrum and at these wavelengths, the solar output is nearly constant (Withbroe, 1989). At higher energies (extreme ultraviolet, X-ray and gamma-ray wavelengths) the solar radiative output becomes more highly variable, as it does also at radio wavelengths.

1.3 Description of Sunspot Behaviour

1.3.1 Historical Perspective

Many references to naked eye observations of the appearance of sunspots on the solar disk exist, dating back to the fourth century B.C. with Theophrastus of Athens (c.370-290 B.C.). More systematic records of naked eye observations belong, however, to the Chinese. These records contain 112 descriptions of sunspots in the official Chinese histories for the period 28 B.C. to A.D.1638. In the west, however, in the literature of the same period, those references to naked eye sunspots which do exist are rare and fragmented, (Bray and Loughhead, 1965), including fourteenth century Russian chronicles of sunspots seen through the haze of forest fires. Bray and Loughhead mention the suggestion that this was partly a consequence of the teachings of Aristotle, who maintained that the Sun was a perfect body without blemish, a belief which became part of the theology of the Orthodox Christian Church in the Middle Ages. This belief lasted up until the age of the telescope with even Kepler (1571-1630) mistakenly attributing the appearance of a sunspot to a transit of Mercury.

Following the invention of the telescope, the real study of solar physics began in 1611 with the first sunspot observations using telescopes. This discovery is credited among four men: Johann Goldsmid (1587-1616) in Holland, Galileo Galilei (1564-1642) in Italy, Christopher Scheiner (1575-1650) in Germany and Thomas Harriot (1560-1621) in England. Of the four, it is uncertain who actually made the first sunspot observation, but Goldsmid, more commonly known as Fabricious, receives the priority of publication. Characteristically though it was Galileo who carried out some of the most thorough early research into sunspots. He disposed of once and for all the suggestion that the spots could be small planets orbiting the Sun by showing that this would be incompatible with the observed changes in their size and shape. Since they must belong to the Sun, regular observation showed that the Sun rotated around a fixed axis with a period roughly equal to a lunar month. He noticed the tendency of spots to group and the relative motion of spots within a single group. Finally, he noted the fact that sunspots were usually confined to two narrow belts, extending around $\pm 30°$ of the equator. In total, these findings constituted the basic knowledge of sunspots for the next 150 years.

The main importance of Scheiner's work was in the length of time over which he carried out his observations. The collection of sunspot drawings which were published in Scheiner's volume *Rosa Ursina sive Sol*, greatly helped later work to trace the behaviour of the sunspot cycle back to the earliest recorded telescopic observations. In comparison to the other three men, very little is known about the life and work of Thomas Harriot.

In 1769, Alexander Wilson, professor of astronomy in the University of Glasgow, studied the physical appearance of a large sunspot as it passed across the solar disk. On approaching the limb, the penumbra on the side remote from the limb appeared to contract until it disappeared. When the spot reappeared at the east limb, the same behaviour was shown by the penumbra on the opposite side of the spot, now the one remote from the limb. Wilson initially attributed this behaviour to the hypothesis that the spots were saucer-shaped depressions on the Suns's surface. Bray and Loughhead (1965) have attributed the true explanation of the Wilson effect to the higher transparency of the spot material compared with the photosphere. In fact these are the same in so far as the surface with opacity $\tau = 1$ is saucer-shaped. Nevertheless, Wilson's work was ahead of its time in that it ranked as the first physical investigations into the properties of an individual sunspot.

1.3.2 The Sunspot Cycle

In 1843, a German amateur astronomer Heinrich Schwabe announced a possible 10- year periodicity in the number of sunspots he had observed over the previous 17 years (Schwabe, 1843). A full publication of his observations in 1851 confirmed his announcement to the scientific community and led to him being awarded the Gold Medal of the Royal Astronomical Society in 1857. For the

last 150 years, the continued existence of an approximate 11-year cycle has been well established and studied as the sunspot cycle. Furthermore, the number of sunspots acts as a convenient index for the level of almost all other activity on the Sun, including active regions, plages, prominences and flares. In this way, the sunspot cycle provides the most easily observed and longest direct source of the history of solar activity, providing a basis for much of the current understanding of the Sun and subsequent solar-terrestrial interactions.

In 1848, Rudolf Wolf of the Zürich Observatory introduced the relative sunspot number R as a measure of spot activity on the Sun. The Wolf Sunspot Number is defined by

$$R = K(10g + f) \tag{1.1}$$

where f is the total number of spots on the visible disk, irrespective of size, and g is the number of spot groups. The coefficient K is a reduction factor which depends on the observer's method of counting spots and sub-dividing them up into groups, on the size of the observer's telescope and the magnification used, and also on the seeing conditions. In this way, it has been possible to scale measurements made from observatories and amateur astronomers around the world to the Zürich scale. Wolf also carried out extensive searches into the past records to see if the sunspot cycle could be reconstructed back as far as the first observations by telescope. He concluded that the 11-year cycle had indeed been in operation since 1700 and probably earlier. A complete compilation of sunspot numbers, including Wolf's reconstructed values, was published by Waldmeier (1961). Four tables are included in Waldmeier's work, covering

(i) the years of minimum and maximum sunspot activity for the period 1610-1960;

(ii) annual mean sunspot numbers for 1700-1960;

(iii) monthly mean sunspot numbers, smoothed and unsmoothed, for 1749- 1960;

(iv) daily sunspot numbers for 1818-1960.

All of these tables have since been continuously updated to the present day. Since around 1850, with the definition of the Wolf Number established, sunspot data are probably very well observed (Eddy, 1977), with little noise in the records. Eddy points out, however, the limitations which exist in the data before this date. Using three categories of reliability, the data can be graded thus: good from 1852 to 1818, fair from 1818 to 1750 and poor from 1749 to 1700. He also suggests that Wolf's sunspot numbers prior to 1749 may represent little more than an extrapolation to fit the 'normal' cycle. Figure 1.1 plots the annual sunspot number from 1700 to the 1989. For the period since monthly mean sunspot numbers have been reconstructed, the sunspot cycle has been numbered consecutively with Cycle 1 starting at the minimum in February, 1755. At the current time, we are on the descent phase of Cycle 22, with the last sunspot maximum in June, 1989.

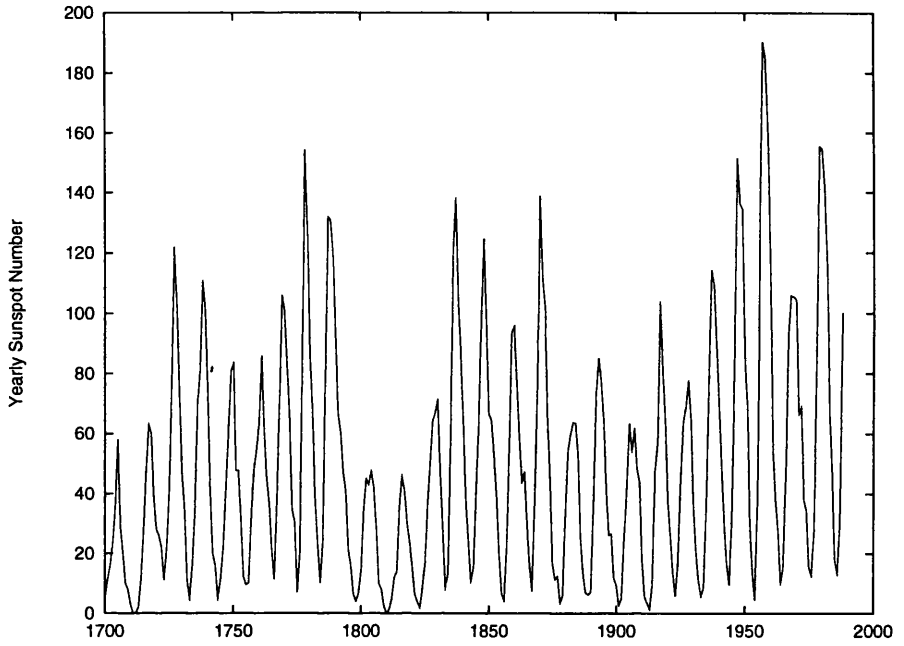


Figure 1.1: Sunspot numbers 1700-1989.

Although the sunspot cycle is often referred to as the 11-year cycle, over the span of 22 cycles, there has been a spread in the actual lengths of individual cycles, ranging from 9 to 13.6 years. The number of spots visible on the solar disk varies considerably from day to day, due to the 27-day solar rotation (since new spots have a tendency to appear close to the positions of old spots) and also to the birth and decay of individual spots and groups. The typical lifetime of a sunspot is less than one solar rotation period although a few large spots will exist for longer than this. These effects are increasingly smoothed out as averages over longer and longer timescales are taken. These averages consist of recording monthly values (which are daily values summed and averaged over a month) and further smoothing is then carried out on the monthly data. This smoothing involves calculation of either a 12 or 13 month running mean. The smoothed data referred to in this thesis is the 13 month smoothed data where the value R_{sm} is given by

$$R_{sm} = (R_6 + R_{-6} + 2\sum R_i)/24 \tag{1.2}$$

where R_6 and R_{-6} are the monthly values 6 months ahead and behind the month of interest and the summation is over the 5 months on either side and including the month of interest. In later sections both the smoothed and unsmoothed data will be analysed. The amplitude of the 11-year cycle is still very variable, however, with a lowest smoothed monthly mean sunspot number at maximum of 48.7 for cycle 6 in 1816 to a highest of 201.3 for cycle 19 in 1958. Table 1.1 lists the values of sunspot minimum and maximum for the 22 numbered cycles and also includes the

Cycle No.	Min. Date	Max. Date	Rmin	Rmax	Asc. (m)	Desc. (m)
1	02 1755	06 1761	8.4	86.5	76	60
2	05 1766	09 1769	11.2	115.8	39	70
3	06 1775	05 1778	7.2	158.5	35	76
4	09 1784	02 1788	9.5	141.2	41	122
5	04 1798	02 1805	3.2	49.2	82	65
6	07 1810	05 1816	0.0	48.7	70	83
7	04 1823	11 1829	0.1	71.7	79	48
8	11 1833	03 1837	7.3	146.9	40	76
9	07 1843	02 1848	10.5	131.6	55	94
10	12 1855	02 1860	3.2	97.9	50	85
11	03 1867	08 1870	5.2	140.5	41	100
12	12 1878	12 1883	2.2	74.6	60	74
13	02 1890	01 1894	5.0	87.9	47	96
14	01 1902	02 1906	2.7	64.2	49	89
15	07 1913	08 1917	1.5	105.4	49	71
16	07 1923	04 1928	5.6	78.1	57	65
17	09 1933	04 1937	3.5	119.2	43	82
18	02 1944	05 1947	7.7	151.8	39	83
19	04 1954	03 1958	3.4	201.3	47	79
20	10 1964	11 1968	9.6	110.6	49	91
21	06 1976	12 1979	12.2	164.5	42	81
22	09 1986	06 1989	12.3	158.5	34	-

Table 1.1: Summary of properties of sunspot cycles 1-22.

ascent and descent times for each cycle. From these times, the asymmetry which exists in the profiles of individual cycles is clear, with cycles with large maxima tending to rise and reach this maximal value more quickly. It is only for small amplitude cycles that the ascent and descent phases are of more equal length. It can be seen that the rise time varies between 34 and 82 months with an average value of 51.3 with corresponding descent values of 48, 122 and 80.5. In talking about sunspot maximum, this refers to a number maximum in the Zürich International Sunspot Number. This generally happens when the Sun has many medium-sized spots on the disk, but not necessarily the largest and more active spots. These tend to occur just before or just after the number maximum. The average sunspot number at maximum for the last 22 cycles is 113.8 although recent solar activity has been higher than in the past with 3 of the 4 largest maxima occurring in the last 4 cycles.

While it took around 200 years from the first telescopic observations of sunspots, to the discovery of the sunspot cycle by Schwabe, it was assumed very quickly that the 11-year cycle of solar activity had always existed. The historical evidence, however, does not necessarily back up this assumption. In the late 1800's, first of all Sporer (1887, 1889) and then Maunder (1890) suggested the existence of a 70 year period in the late seventeenth and early eighteenth centuries when virtually no sunspots were recorded. During this period (1645-1715), which is commonly called the Maunder Minimum, the level of activity at no time appeared to rise above a value of $R = 5$, less than the minima of

many of the recently recorded cycles. At such low values of R, it is not possible to confirm that the 11-yr cycle was operating at all during the Maunder Minimum. Also, although sunspot activity was definitely higher in the years preceding 1645, the observations are scarce enough to cast doubt on the very existence of the cycle pre-Maunder Minimum. Related work by Eddy et al. (1976) has suggested that the solar rotation was different at the start of the Maunder Minimum from solar rotation in the modern era, with rotation at the equator about 3% faster than now and differential rotation enhanced by a factor of 3. At the very least this would suggest that the nature of solar activity may have changed in modern times.

The natural reaction to this discovery is to look for any observations made prior to the Maunder Minimum in an effort to determine the early history of solar activity. Eddy (1977) details 5 factors which could shed light on this question. These are naked-eye sunspot observations, auroral records, the appearance of the corona at eclipse, tree rings and Carbon 14 records. The first two of these are more useful when used together to establish any trends in the history of the Sun. This is because in general the paucity of observations before the age of the 'enlightenment of science' around the late 16th and early 17th centuries does not allow any firm conclusions to be drawn. Long term averages of these observations support each other in suggesting that the 12th century was a period of relatively high solar activity while another minimum, similar to the Maunder period, was in existence between 1450 and 1540. In terms of an 11-yr cycle, there is nothing in these historical records which can back up the existence of such a cycle before the age of the telescope.

The form and extent of the solar corona evolves in a well-known way with changing levels of solar activity. Eddy (1976) confirmed reports that the corona may have been in abeyance during the Maunder minimum, since eclipse reports during that period seem to suggest a corona which was much reduced in brightness. Unfortunately there are no unambigous descriptions of a structured corona in the historical records before 1645 (Eddy, 1976) and so it is not possible to use this factor in any positive way. Douglass (1919) raised early hopes that evidence of past solar cycles could be found in patterns of the rings in trees but these have subsequently been dashed. Instead this work has proved of much use in archeaological and local climate research but since no link between local climate and solar activity has been firmly established, the use of tree ring studies has been discounted.

Finally the terrestrial record of Carbon 14 as an index of solar activity has been investigated by several authors (see Eddy, 1977 and references therein). Carbon 14 is formed through interactions of galactic cosmic rays with atmospheric nitrogen. Cosmic rays are modulated by changes in overall solar activity. Thus the relationship with sunspot number is inverse: at times of low solar activity, the shielding of the Earth by the Sun's extended field is less and so more cosmic rays interact with the atmosphere and Carbon 14 production increases. The situation is reversed at times of high solar activity. There are many other terrestrial effects which affect the ratio of C^{14}/C^{12} but it is

possible to look at the deviations in the record and confirm the presence of the major features of solar history which have already been described, as well as a suggestion of possible earlier changes of equal magnitude.

Thus although the very early history of solar activity cannot be analysed in anything like the same detail as provided for the last 300 years or so, it is clear that definite features exist in solar activity levels. What is undeniable, however, is the existence of the solar cycle variations since the end of the Maunder Minimum. The effect of these variations and the importance of understanding them is examined in a subsequent section (1.5). Before that, for the sake of completeness, a summary of the physical characteristics of sunspots is provided.

1.3.3 Physical Characteristics of Sunspots

Although the main body of this work consists of numerical methods for analysing the variations in the level of activity observed on the Sun, it is useful to provide a more physical description of the phenomena under study. The following description of the characteristic behaviour of individual sunspots, and also groups, will allow a full picture of the sheer complexity of this aspect of the solar activity cycle to be painted. Sunspots represent the most intense phase of an active region in the solar atmosphere. They are observed in the photosphere as dark spots which are cooler then their surroundings with exceptionally strong magnetic flux, typically 2-3kG. The central dark area of the spot is called the *umbra* with a typical diameter of 10-20,000km. The effective temperature of the umbra is around 3700K, which is considerably less than the surrounding photospheric temperature, 5800K.

Although sunspots are concentrated within two bands $\pm 30°$ from the equator and rarely appear higher than this, at any one time there can be a great spread in latitude of visible sunspots. Carrington noted in 1858 a dependence on the average latitude of visible sunspots and groups on the phase of the sunspot cycle, with sunspots early in the cycle tending to appear at higher latitudes. Later work by Wolf and Sporer confirmed this dependence which is often referred to as Sporer's Law. It is best demonstrated, though, as the *butterfly diagram* of Maunder (1922) (see Figure 1.2). This effect represents a drift in *locus* of the spots – individual spots exist for less than one solar rotation period in general and show little or no latitude motion relative to the photosphere. At the onset of a cycle, the average spot latitude is 28°, six years into a cycle this has reduced to 12° and after 11 years, it is 7°. At this point some overlap usually occurs with spots of the new cycle appearing at higher latitudes.

Between 1908 and 1925, daily observations of the magnetic polarities of sunspot groups using the Zeeman effect (Hale, 1912) were carried out at the Mount Wilson Observatory. Over 2200 sunspot groups were studied and led to the publication of two important works (Hale et al., 1919, Hale and Nicholson, 1925) in which the authors introduced the Mt Wilson magnetic classification of sunspot

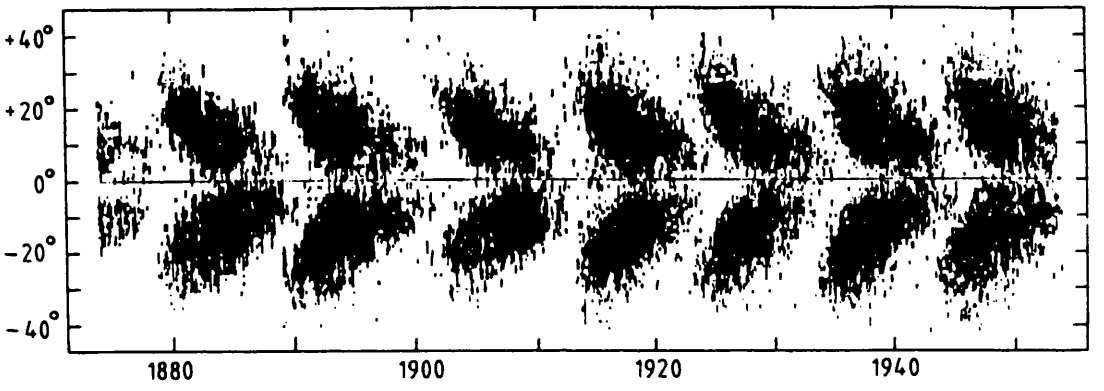


Figure 1.2: Maunder's butterfly diagram showing the latitude drift of sunspots during the 11-year solar cycle.

groups and formulated the Hale-Nicholson Laws of Sunspot Polarity. The polarity laws can be summarised as follows. The majority of sunspots ($\sim$ 60%) appear as part of what are classified *bipolar* groups, where spots of both magnetic polarities exist. It is observed that for all bipolar groups in the same solar hemisphere, the same polarity leads while the following spots in the group are of the opposite polarity. In the other hemisphere the polarities of the leading and following spots are reversed. It is also observed that the polarity of the leading spots in a bipolar group reverses for each new sunspot cycle. The rest of the sunspots usually fall into the *unipolar* classification although most of these spots are followed by "invisible spots" , areas of opposite magnetic polarity which have not formed sunspots or have been left following the decay of previously existing spots. Around 1% of groups are classified *complex* meaning that the distribution of polarity amongst their spots does not fall into the earlier classes. In many of these cases, complex groups are associated with highly active regions producing many flares.

Finally, another feature of sunspots ("Joy's Law") is the observed fact that the leading spots in most groups are angled closer to the equator than the following spots. Thus the magnetic axis shows an average tilt of 5.6° to the E-W line. Hale also searched for a general solar field for many years and made claims to have observed one. Babcock and Babcock (1955), however, used the magnetograph to confirm the existence of a polar field, but about ten times weaker than Hale claimed. Babcock (1959) showed that the field reversed around the time of solar sunspot maximum and thus the Sun requires two 11-year cycles to return to the same magnetic configuration. This resulting 22-year cycle is called the Hale Cycle.

1.4 Description of Solar Irradiance Variations

This section discusses the variations of the whole solar irradiance, or total photon output. An extensive and thorough review of this topic was carried out recently (Lean, 1991). This thesis attempts to cover the main points of this review and only pays particular attention to the specific case of the solar 10.7cm flux which will have a more prominent role in the later chapters. The geophysical implications of these variations are discussed in the upcoming section (1.5) covering the solar-terrestrial interactions. Classically the luminosity of a star is one of its principal classifying parameters. Solar luminosity, L in Watts, is related to the diameter, d (metres), and irradiance, S (Watts/m^2), by

$$L = 4\pi(d/2)^2 S \tag{1.3}$$

The launch of the Nimbus 7 satellite and the Solar Maximum Mission carrying respectively the Earth Radiation Budget (ERB) and Active Cavity Radiometer Irradiance Monitor (ACRIM) radiometers allowed simultaneous space-borne measurements of the Sun's total irradiance to be made almost daily between March 1980 and July 1989. More recent measurements have been made a few times a month through radiometers onboard the Earth Radiation Budget Satellite and the NOAA-9 satellite. The accuracy and precision of these results have surpassed a century of ground-based observations and constrained the magnitude of the solar minimum total irradiance to 1367 ± 3 Watt/m^2 and revealed an amplitude of approximately 0.1%. Prior to these satellite observations the constraint on solar variablility was only $< 1\%$ and more importantly could not rule out the possibility that the total irradiance was actually invariant. Evident in the satellite data is a long-term downward trend concurrent with the decline in solar activity following the maximum of cycle 21 in 1979 until the minimum in 1986 was reached. Thereafter an upturn in irradiance values is seen. Although the precision of these results is generally excellent, instrumental problems have still placed some limitations on the ultimate information which can be obtained. In particular there are some differences between the ERB and ACRIM datasets and revisions in the data associated with instrument calibration, pointing, temperature and geometry are still ongoing (Hoyt and Kyle, 1990) in the hope of reconciling these differences.

The first concurrent observations by ACRIM and ERB showing reductions of the solar irradiance by a few tenths of a percent over time scales of weeks, led to the connection being made between these dips and the presence of sunspots on the solar disk (Hudson et al., 1982). Ground based calculations of the total solar irradiance are made by adding up the contributions from quiet and active Sun regions, since the latter affect the solar irradiance. Sunspots suppress photon emission and calculations involving measurement of the areas, locations and contrast of the sunspots verify that large irradiance dips are generated by the passage of sunspots across the face of the Sun as seen at the Earth, in concert with the Sun's 27-day rotation (Hudson et al., 1982, Sofia et al., 1982). The question of how sunpots apparently suppress the solar radiance remains open. If the energy

is genuinely 'blocked', ie trapped inside the Sun, there is a problem as to where it accumulates. The actual energy production cannot be affected as the photon diffusion time is $10^3 - 10^4$ years. One possibility is that the photons are emitted in different directions in the presence of a spot, or perhaps from the back of the Sun.

Sunspots are not the only source of total irradiance variations. Solar faculae represent regions of enhanced emission and so increase the total irradiance, in opposition to the sunspot blocking effect (see, for example, Foukal and Lean, 1986). It has been suggested (Chapman, 1984, Lawrence and Chapman, 1990) that over the lifetime of an active region, the energy blocked by sunspots is equivalent, within $\pm 20 - 30\%$, to that radiated by the faculae. This is possible because although the bolometric contrast of faculae is much smaller than sunspots, in general these regions cover an area about an order of magnitude larger and exist for longer than individual sunspots. This still requires mechanisms to explain the energy balance and storage within an active region and Schatten and Sofia (1987) and Foukal et al. (1983) have suggested possible answers.

The above discussion has concentrated on the behaviour of the total solar irradiance. The picture is different when spectral considerations are made. The Sun's photon output is typical of a black-body radiator at $\approx$ 5770K although the whole spectrum is modulated by solar activity. The total solar irradiance variations of $\leq 0.1\%$ are typical of the visible spectrum emissions which dominate the total solar output. A survey of the Sun's spectral irradiance with an estimate of its variability during solar cycle 21 is reproduced from Lean (1991) in Figure 1.3. The largest amplitude variations of more than an order of magnitude are found in the wings of the black-body spectrum, both at the shortest (extreme UV and X-ray) and longest (radio) wavelengths. The photon emission at the ultra-violet end of the spectrum ($<$ 300nm) constitutes only about 1% of the Sun's total output. It is, however, of considerable interest to both the solar and geophysical communities. Since these emissions are formed over a wide range of heights in the solar atmosphere, they provide an important diagnostic of certain regions of the atmosphere not accessed by any other wavelengths. Also, because the variations are of an order of magnitude greater than those in the longer wavelength visible spectrum, they are consequently an important contributor to the total solar irradiance variability irrespective of the minor role these emissions play in the total solar irradiance. For the geophysical case, solar UV radiation is almost totally absorbed in the Earth's atmosphere and thus has the potential for considerable effect on the terrestrial environment. Before these effects are considered, we shall describe in more detail the database of solar 10.7cm flux. This is a radio wavelength where solar irradiance variations are again more significant than in the visible spectrum.

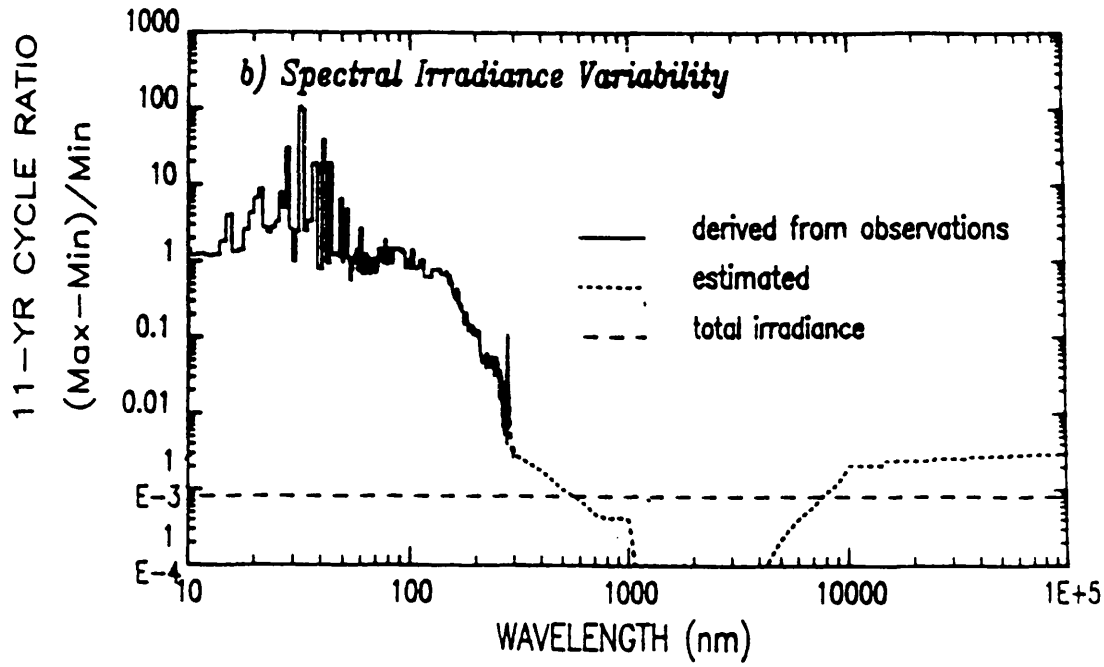


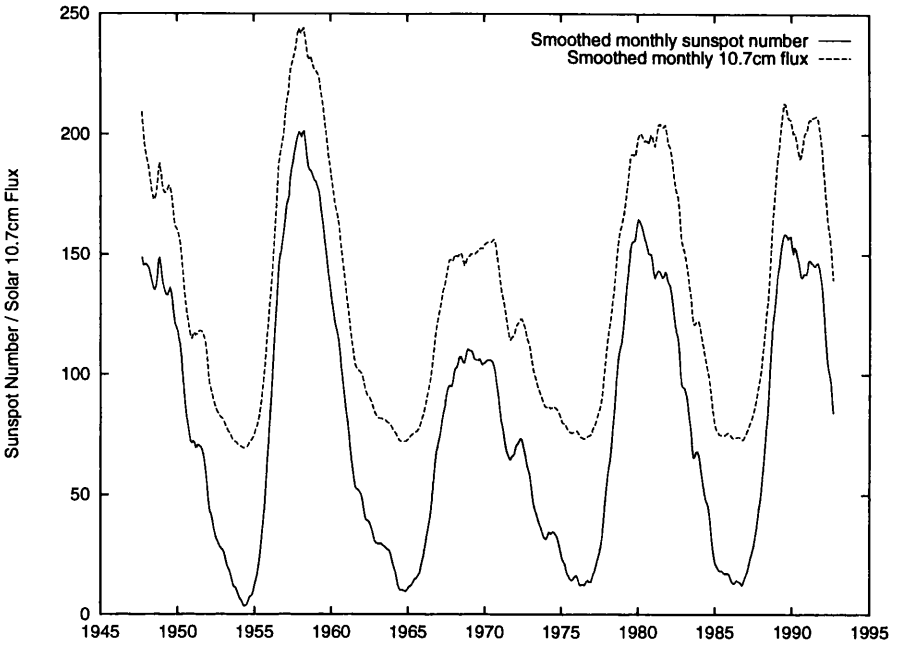
Figure 1.3: Approximate amplitude of the Sun's spectral irradiance variation from the maximum to the minimum of the Sun's 11-year activity cycle. The variations at $\lambda < 300$nm (solid line) were derived from satellite observations during solar cycle 21. The variations at longer wavelengths (dotted line) were determined from knowledge of the solar cycle variation in the fraction of the Sun's disk covered with active regions and of their contrasts. The dashed line indicates the variation during solar cycle 21 of the spectrally integrated (total) irradiance.

1.4.1 Solar 10.7cm Flux Measurements

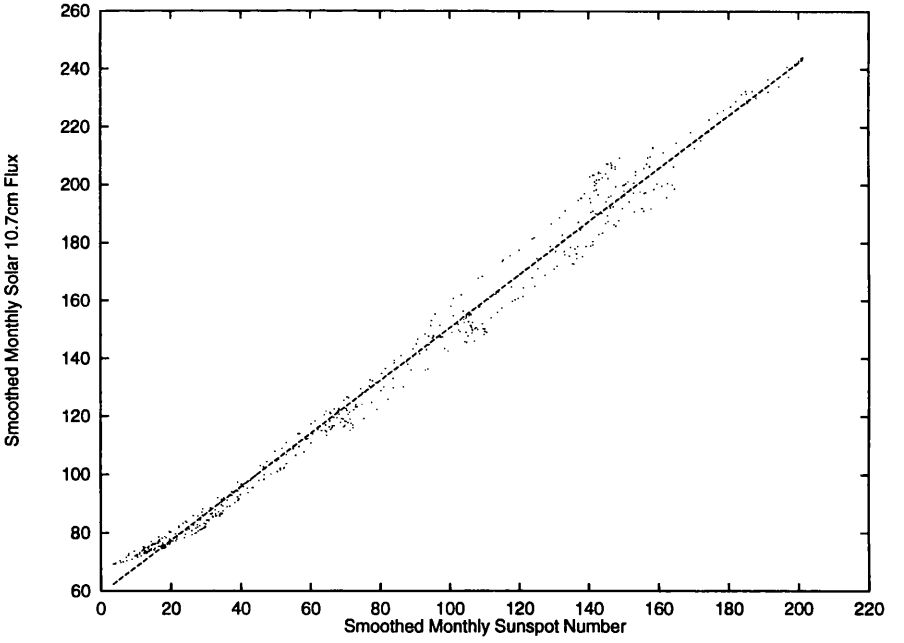
Scientific measurements of the solar flux at the 10.7cm (radio) wavelength have been made daily since 1947. This wavelength was chosen primarily because the Earth's atmosphere is transparent to this radiation and so measurements could be made at the Earth's surface, without the need for spaceborne equipment. Since the dawning of the space age, more consistent measures of other wavelengths have been made. The F10.7cm flux remains important due to the consistency of its measurements and also, as we shall see, through the correlated behaviour with radiation at other wavelengths (for example, ultra-violet) which would require measurement above the atmosphere. The behaviour of solar activity has already been established through consideration of the sunspot cycle (Section 1.3.2). It is an obvious place to start to compare the measurements of the solar flux against the sunspot cycle for the overlapping years since 1947. This is done in Figure 1.4(a) and it is clearly seen that a strong correlation is present. A scatter plot of the two indices over the same period (Figure 1.4(b)) and a least squares regression on the smoothed monthly data available from 1947 to the start of 1993 give a best fit line with equation

$$F = 0.915 * R + 59.16 \tag{1.4}$$

where F and R are the smoothed monthly solar 10.7cm flux and sunspot number respectively. The close association between these indices allows for the solar flux values to be extrapolated into the



(a)



(b)

Figure 1.4: Figure (a) plots the smoothed monthly sunspot number along with the solar 10.7cm flux for the period 1947-1993. Figure (b) shows a scatter plot of the two indices over the same period, with the best fit straight line calculated by least squares.

past to build up a bigger database for its past behaviour. This proves useful and necessary in many prediction methods currently in use. This extrapolation is of course based on the assumption that the solar flux variations have always been so modulated with the sunspot cycle.
The correlation with extreme UV radiation (EUV) is also of great significance. The EUV flux has generally only been measured over relatively short intervals but in these intervals, the behaviour of the 10.7cm flux is a good indicator for its behaviour. Since the EUV flux plays a primary role in heating the outer atmosphere of the Earth, the 10.7cm flux has consequently been used as a proxy for the EUV flux and included in atmospheric calculation models (Hedin and Mayr, 1987).

1.5 Influence on the Space Environment

Up until now we have discussed solar activity behaviour as a purely phenomenological subject. This has detailed the features which are observed on the Sun and the radiation which is emitted from it. While these variations are interesting enough in themselves to warrant further study, they do also have consequences for the near-Earth space environment. This makes the understanding of these variations as much of a necessity as anything else. This section will cover four main subject areas. The first of these will explain briefly how solar variations can interfere with the terrestrial system and will be contained in the remainder of these opening comments. The other three will be in the form of subsections concentrating on some of the specific consequences of this inteference, namely satellite drag, solar particle events and geomagnetic storms. These last three topics will enable some specific details of solar-terrestrial interactions to be explained in more detail.
The effects of solar activity on the near-Earth environment can be described in terms of direct and indirect effects (Gorney, 1990). The former are classified by rapid changes in the solar UV and X ray illumination of the Earth's atmosphere and ionosphere, particularly during solar flares. Indirect effects are caused by more complex interactions between the solar wind and the coupled magnetosphere-ionosphere-atmosphere system. A full description of some of the mechanisms currently thought to explain the indirect coupling of solar activity with the Earth's ionosphere and magnetosphere is beyond the scope of this thesis. The reader is directed to a recent review article on this topic by Gorney (1989) as well as the paper by Gorney quoted above. Instead the upcoming subsection on geomagnetic storms will describe some of the causes and consequences of magnetospheric interactions. The direct coupling mechanisms will also be covered in more detail through the use of 'case studies', namely solar particles events and satellite drag.

1.5.1 Solar Particle Events

These events constitute one of the most direct interactions of solar activity on the near-Earth space environment. Solar flares give rise to the sudden release of very large amounts of energy in the

solar chromosphere and corona. Associated with these can be the emission of very high energy accelerated particles which, in the right conditions, can interfere with the near-Earth environment. These events are fairly infrequent (<16 per year, each year for the last three solar cycles, Smart and Shea, 1989) but with energies of 10MeV to 1GeV are the most energetic tangible manifestations of solar activity and can have severe consequences for man-made systems operating in space. The propagation of solar particles depends on the exact energies of the particles emitted. Very high energy particles can intercept with the Earth after $\sim$ 8 minutes while lower energy particles will take a few hours. Since the particles follow Archimedian spiral trajectories linked to the interplanetary magnetic field topology, the onset time and intensity of an event is strongly dependent on the helio-longitude of the originating flare.

The space hazards posed by solar particle events are considerable. During one solar particle event, an unshielded human would receive a radiation dose equivalent which is three orders of magnitude greater than the annual limit established by governments for the general population. There is a particular danger for extended space missions, such as a Mars mission, as this would be completely outwith the protection which can be provided by the Earth's magnetosphere. As well as the human risk, these high energy particles carry a high element of danger for sophisticated electronics on board satellites. There is also a cumulative detrimental effect on the solar panels which provide the source of power for many current satellite missions. The priority of these effects are evidenced by the recent Eureca mission devoted to studying this problem.

In terms of the prediction of solar particle events, much of the current research depends on a better understanding of the solar flare mechanism, and thus the instances when particle emission occurs. On time scales greater than a few days, the prediction problem becomes one of understanding the future development of active regions into areas where flares are likely (Kunches et al., 1991). Although as will be seen, this thesis does not directly concern itself with the predictive ability of neural networks in this precise field, in general the prediction of levels of solar activity through sunspots (and associated active regions) provides some input to this problem and establishes the methodology which could subsequently be applied to this more specialised area.

1.5.2 Implications for Satellite Drag

As early as the introduction to this chapter, the importance of solar radiation as the energy source for the Earth was mentioned. On a clear day, the spectrum of the Sun's radiation between 300nm and 1000nm, near the peak of the black-body spectrum reaches the Earth's surface with little atmospheric attenuation. Ultraviolet radiation at wavelengths <300nm is absorbed by atomic oxygen, molecular oxygen, molecular nitrogen and ozone which constitute most of the upper atmosphere. The variability of this part of the solar spectrum has already been discussed (Section 1.4), and there are very clear indications that these variations are geo-effective. Although any direct link

with the lowest levels where weather and climate are experienced is unproven, one principal concern regards the demonstrated influence of solar activity on the Earth's upper atmosphere. Absorption of short wavelength radiation in the thermosphere affects the heating and composition of this layer with the subsequent effect of enhanced atmospheric drag being felt by satellites. This leads to the possibility of orbit decay and re-entry. Indeed, studies of these effects (Walterscheid, 1989), suggest that a satellite at an initial orbit 500km above the Earth would have an orbital lifetime of 30 years under typical solar cycle minimum conditions. This would reduce to only about 3 years for solar maximum conditions, depending obviously on the precise size of the cycle maximum.
Such a considerable variation in orbital lifetime can be understood through consideration of the drag which satellites experience as they move relative to the bulk wind motions of the thermosphere. This deceleration is given by

$$F_D = \frac{1}{2} C_D A \rho V_r^2 \tag{1.5}$$

$$V_r = V_s - V \tag{1.6}$$

where C_D is the drag coefficient, A is the area-to-mass ratio of the satellite, ρ is the total mass density of air, V is the velocity of the bulk winds in the thermosphere, V_s is the velocity of the satellite in a corotating frame and thus V_r is the velocity of the satellite with respect to the bulk wind vector. The greatest effect on F_D due to solar cycle variations is through the dependence on the air density since, above $\sim$ 300km, ρ at a given point can vary by an order of magnitude due to solar cycle variations (Walterscheid, 1989). This is because $\rho(z)$ is exponential in behaviour and so $\delta\rho/\rho \approx \delta z$. Also, since $V_s >> V$ the F_D dependence on variations in V is slight as is the dependence on the drag coefficient even although C_D can show some variation due to changes in composition and temperature of the atmosphere.
The precise mechanisms behind density changes due to solar cycle effects are detailed in Walterscheid (1989) and references therein. In essence, the variations are due to temperature changes and temperature-induced composition changes which affect the ratio T/m, where T is the temperature and m the mean molecular weight of the atmosphere. This ratio is equivalent to the exponential scale height which gives the dependence of density at height z in the atmosphere. Walterscheid (1989) also investigates the effect of solar cycle variations on geomagnetic activity (see Sections 1.5.3 and 5.4.1 for more details) and concludes that they are not a major influence on satellite lifetimes on the solar cycle timescale of variation. The short term variation (days to weeks) will have consequences for tracking and cataloguing of space objects.
The effect on satellite orbits is certainly one of the main reasons why the prediction of the solar activity cycle remains so important. For example, the Solar Maximum Mission satellite underwent rapid orbit decay at the end of the 1980's due to the rapid early rise of solar cycle 22.

1.5.3 Geomagnetic Storms

Since the last century, it has been recognised that some relation exists between certain solar events and strong magnetospheric disturbances. These strong disturbances are called geomagnetic storms. With the discovery of the solar wind and the shock waves which propagate with it, storms were associated with the Earth passage of shock disturbances (Hundhausen, 1972). The debate has continued since then over the solar origin of the interplanetary shocks which give rise to geomagnetic storms. Initially solar flares were thought of as the prime cause but the association has turned out to be unsatisfactory since many shocks have been observed without flares and vice versa. Joselyn and McIntosh (1981) and Wright and McNamara (1983) discussed the sudden eruption of solar prominences as likely sources of shocks but again these events have been observed to occur independently of each other.

Coronal Mass Ejections (CMEs) are now considered by many authors to be the solar phenomena which produce interplanetary shocks. Correlated studies of coronal and interplanetary shocks (for example Sheeley Jr. et al., 1985) showed that almost every interplanetary shock observed was related to a CME. There are still problems with this association, however, since most CME's are not related to shocks and so a selective process must operate. Bravo and Rivera (1994) has covered some of the recent debate on this subject and proposed the scenario where interplanetary shocks are related to regions on the Sun containing both a coronal hole and a flare or an erupting prominence.

The interaction becomes more complicated at the Earth since not all interplanetary shocks are effective in producing geomagnetic storms. Thus a coupling between the solar wind and the terrestrial magnetosphere is required and exhaustive studies have shown this coupling to be magnetic, as first proposed by Dungey (1961). Gonzalez and Tsurutani (1987) and Gosling et al. (1991) have carried out intensive studies of geomagnetic storms and related them to solar wind parameters and a southward oriented component of the interplanetary magnetic field. Geomagnetic storms themselves have been classified into two types, storm sudden commencements, which begin very rapidly, and those which begin gradually. Mayaud (1975) catalogued a hundred years of storm sudden commencements. The different types of magnetic storm, although they clearly exist, are not distinguished by use of geomagnetic indices, although severe storms are usually sudden in onset. A discussion of geomagnetic activity indices is delayed until Section 5.4.1 when the time series data is analysed.

Finally in this section, the effect of geomagnetic storms on the Earth is discussed briefly. One of the main effects is that of geomagnetically induced currents (GIC). Disturbances in the magnetic field induce electric fields at the Earth. These electric fields produce currents in man-made conductors, most notably in high-voltage power systems. Boteler (1993), Viljanen (1993) and Lundstedt (1993) discuss the importance of being able to predict the occurrence of GIC's in the power systems of

Canada, Finland and Sweden respectively. All of these papers have discussed the potential use of neural network or artificial intelligence techniques to predict these events. An extreme example of a GIC was the occurrence of a power outage in eastern Canada in March 1989 which affected approximately 6 million people for over 9 hours. This was caused by geomagnetic events and speaks for the practical importance of these phenomena. Other effects are also noticeable, including corrosion of pipelines, malfunctions of instruments used for geological surveys, tracking problems of satellites (see also Section 1.5.2) and radio and telecommunication problems.

1.6 Established Prediction Methods for Solar Activity

The aim of this section is to summarise many of the prediction methods, algorithms and techniques which have been suggested and utilised in making predictions of future levels of solar activity. The number of methods in the literature are too numerous to make a complete survey realistically possible. A recent review was carried out by Withbroe (1989) in which the main types of technique were categorised with example methods of each described. This section draws heavily on this review as a starting point although has the slight advantage of being written post cycle 22 maximum, which the original paper was not. Solar cycle 22 was of immediate and particular interest from a very early stage after its onset in September, 1986. This was due to an extremely sharp onset during the first couple of years. From Table 1.1 and an analysis of ascent slopes and corresponding maxima (eg Wilson et al., 1986), it is clear that higher maxima have a tendency to follow sharp initial increases in the sunspot number. This may simply be due to the fact that the maximum of a cycle is, in general, reached before the halfway point in a cycle and so for higher values to be obtained within a similar time since onset, this necessitates a more rapid climb. Thus most of the excitement was based on the possibility of cycle 22 beating the maximum smoothed value of 201.3 for solar cycle 19 to become the highest maximum ever (see for example, Brown, 1988 and Schatten and Sofia, 1987). The rise stage of the last four cycles is shown in Figure 1.5 to show the rapid onset of cycle 22. As seen, for the first two years, it was looking likely to rival the maximum of cycle 19 until the actual maximum was reached at a value of 158.5. This would have had the mixed blessings of high levels of solar activity and all the associated effects under study by spaceborne instrumentation for the first time, but with the expected side effects which high activity has on this equipment (see Section 1.5.1).

Due to the lack of a complete quantitative theory for solar-stellar activity cycles, the only way to predict the future behaviour is from a study of the past history. The problem regarding this approach is whether enough information regarding all the modulations which may affect each individual cycle is contained in the limited reliable data which is available. Most of the effort in future predictions has been concentrated on providing the maximum value of the current or next

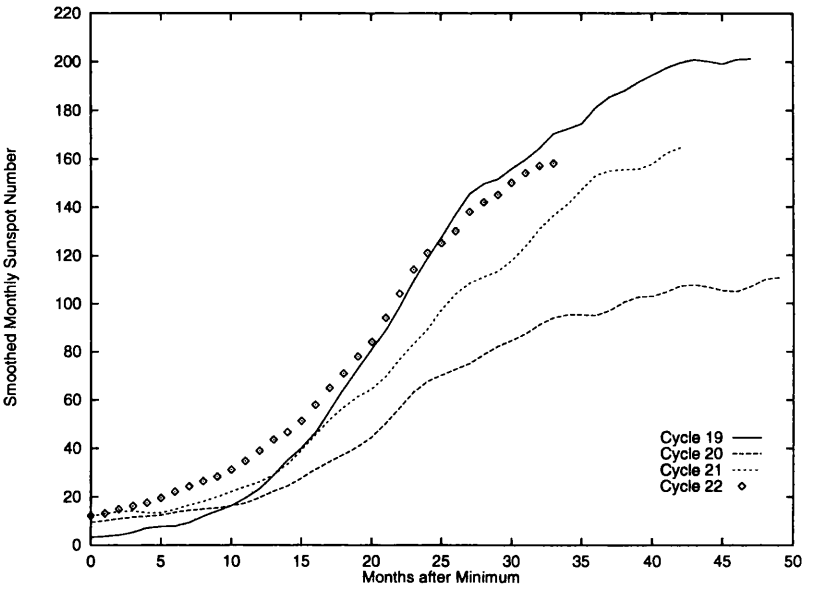


Figure 1.5: The onset stages of the last four solar cycles, showing the rapid rise of cycle 22 which gave rise to excitement that this cycle would reach record levels of activity.

cycle. Withbroe (1989) lists thirty-nine different predictions which were made for the maximum value of solar cycle 21. These ranged over smoothed monthly values from 30 to 200 with an actual maximum of 160. Of these methods, 22 came within ±20% but only 8 were within ±10%. Accuracy of any predictions of the date of the maximum was not considered.

The zeroth order approach to prediction considers each solar cycle as independent of the rest. Then the probability of the actual maximum sunspot number R being greater than R_{max} can be estimated from historical data based on the observed mean value of $\overline{R}_{max} = 116.4$, with standard deviation 41.2. These values are calculated by Withbroe using only the observed data after 1850 which correspond to cycles 10-21. Table 1.2 gives these probabilities based on the normal distribution. The zero-order approach ignores other information which is in existence through sunspot observations. In particular it takes no account of the fact that a correlation exists between the sunspot number at times t_1 and t_2 in a cycle. This is evidenced by the already mentioned tendency of large solar maxima to be preceded by a rapid early rise. It also ignores the suggestion (reported by Hirman et al., 1988) that even and odd cycles behave differently. This could potentially be related to the 22-yr magnetic cycle (Section 1.3.3)

Withbroe quotes a further sixteen predictions which had been made for cycle 22 maximum by the time of writing the paper (see Table 1.3). He classifies them into three different class of prediction method: statistical, precursor and the method of McNish and Lincoln. Statistical methods rely on determining trends in the sunspot record to establish how the next cycle will behave. Inclusion of

R_{max}	$\% > R_{max}$	$\% > R_{max}$ (normal distribution)
200	8	2.1
180	8	6.1
160	17	14.5
140	33	28.1
120	33	46.4
100	58	65.5
80	75	81.1
60	100	91.5

Table 1.2: Based on sunspot data from cycles 10-21, the observed percentage of occurrence for maximum spot number $> R_{max}$ (reproduced from Withbroe, 1989).

22-year variations which result in differently shaped even and odd cycles and also the possibility of longer term modulations of the order of 80-100 years (Gleissberg, 1958) are examples of the trends which are included in these models. In the latter case, the existence of only 200 years of reliable data leads to obvious difficulties in determining accurately any periodicities in the sunspot cycle on this timescale. At this point the ability to extend the sunspot record using other proxy indices could potentially prove fruitful but as explained in Section 1.3.2 the uncertainties arising from these proxy indices do not provide a reliable basis for this extension.

Withbroe (1989) classifies a slightly different type of prediction technique as 'precursor' methods. The basic premise of this class of methods is that the behaviour of the solar magnetic field in cycle N determines the conditions in cycle N+1. An example of this is the correlation of some geomagnetic index during the previous cycle (or at minimum) with the magnitude of sunspot number at the following maximum (for example Schatten et al., 1978). Correlations may also be derived between certain selected solar features in cycle N with an inference drawn as to the size of the following maximum. Out of the 39 techniques for cycle 21 predictions, 12 were classified as precursor methods and these were in general more successful in this first test.

A table of predictions for the maximum of cycle 22 is reproduced from Withbroe in Table 1.3. The specific details about the different statistical or precursor methods are contained in the individual references listed. From these different predictions it is clear that the statistical methods all consistently predict lower values for the maximum value. As a general class they also underestimated the size of cycle 21. The reality of the behaviour of cycle 22 was that after the initial rapid rise in the first two years it reached an early maximum of 158.5 by June, 1989, where it levelled off for about a year before entering the descent phase which it is currently on. This was a surprise to many observers, representing a deviation from the originally expected extreme maximum, perhaps explaining the many over-estimates to be found in Table 1.3. As a group the precursor methods did actually provide a good estimate since, assuming equal weighting to them all, the mean value was 154 ± 45. The question remains though as to which individual method to use.

Author	Technique	Cycle 22 Prediction	Error, 2σ
Brown (1988)	Precursor	174	35
Brown (1988)	Precursor	175	35
Kane (1987)	Precursor	165	35
Lantos and Simon (1987)	Precursor	110	40
Sargent (1978)	Precursor	119	50
Schatten and Sofia (1987)	Precursor	170	50
Thompson (1987)	Precursor	163	none
Wilson (1988a)	Precursor	145	8
Wilson (1988c)	Precursor	154	70
Wilson (1988c)	Precursor	164	70
Lindberg (1989)	McNish-Lincoln	179	-
NOAA/NESDIS	McNish-Lincoln	192	51
NOAA/SEL	McNish-Lincoln	203	52
Brown and Simon (1986)	Statistical	100	45
Wilson (1984)	Statistical	107	73
Wilson (1988b)	Statistical	75	49

Table 1.3: Predictions for maximum smoothed monthly sunspot number for solar cycle 22.

Most of the methods described under the general class headings described above concentrate on predicting solely the value of the next solar maximum. The final 'class' of prediction method which Withbroe includes is that of the McNish and Lincoln method. It is a long standing method, (McNish and Lincoln, 1949), which has often been used to generate predictions of the sunspot number month by month. The method is statistical in nature and was originally applied to predicting the annual sunspot number. The method has been altered and modified to enable more accurate predictions of monthly data. It, and variations, have been used by the National Oceanic and Atmospheric Administration (eg Hirman, 1989), NASA (for example Holland and Vaughan, 1984) and the European Space Operations Centre (for example Mugellesi and Kerridge, 1991) to calculate future values of the sunspot number or the solar 10.7cm flux. Since it has been so widely used and provides predictions not only of solar maximum but also of values at any other time in the cycle, this method is chosen as an ideal benchmark technique against which the predictions afforded by neural networks can be tested. For this reason, it is necessary to describe the method in full detail.

1.6.1 The Method of McNish and Lincoln

In the original paper (McNish and Lincoln, 1949) the authors describe the basis of the method through the following assumptions.

(i) In a time series exhibiting cyclic tendencies an estimate, to a first approximation, of a future value of the series is the mean of all past values for the same part of the cycle and

(ii) this estimate can be improved by adding to the mean a correction proportional to the departures of earlier values of the same cycle from their respective means, the factors of proportionality being determined by the method of least squares.

This work was concerned mainly with the prediction of smoothed annual sunspot numbers although consideration of using three-month means was also mentioned. Holland and Vaughan (1984) adapted this method with the aim of improving its applicability to predicting (smoothed) monthly mean values. The main area of change necessary for working with monthly numbers is the ability to take account of the different lengths of cycles when calculating the mean cycle and subsequently basing predictions from it. Following this approach, the particular method of implementation is demonstrated by Mugellesi and Kerridge (1991). In this way,

- the mean period (P) of previous sunspot cycles is calculated to the nearest month.
- the individual cycles are then resampled at P+1 points, so identifying the same phase of each cycle. This requires interpolation between actual data points.
- a mean cycle is calculated using the resampled sunspot values.
- the departure from the mean cycle for month m of cycle c, $D_m(c)$, is forecast using

$$D_m(c) = k_1 D_{m-1}(c) + k_2 D_{m-2}(c) + \ldots \tag{1.7}$$

where D_{m-j} is the observed departure j months ago in the current cycle. It has been found that Equation 1.7 can be truncated after the first term with no noticeable effect on prediction accuracy. In this case the coefficient k_1 is determined by the method of least squares. Thus,

$$k_1 = \sum_{i=1}^{N} D_m(c-i) D_{m-1}(c-i) \; / \; \sum_{i=1}^{N} (D_{m-1}(c-i))^2 \tag{1.8}$$

where N is the number of previous cycles included in the analysis.

This technique has been found to be most useful when predicting relatively short timescales up to about 12 months ahead. The nature of this method results in the prediction uncertainty decreasing with decreasing time to maximum predictions. Withbroe (1989) thus concludes that ultimately it should provide the most reliable prediction although it is important that the 'predict-ahead-time' is still large enough to be useful. For monthly predictions further ahead than about one year, the predicted values tend to relax towards the mean cycle values and so are much less reliable. In this thesis, the method of implementation contained in the *bulleted* descriptions was followed. The software written is described in Section 3.4 although the procedure is identical to the theory contained here. The mean period of the previous solar cycles 1-19 was found to be 133 months, or 11.1 years. This left a prediction set covering cycles 20, 21 and the start of 22 which would be identical to that which was used for neural network predictions. The same phase of each cycle was calculated using linear interpolation between points, although it would be straightforward to extend this to a cubic spline interpolation if necessary. Subsequent testing of the code against

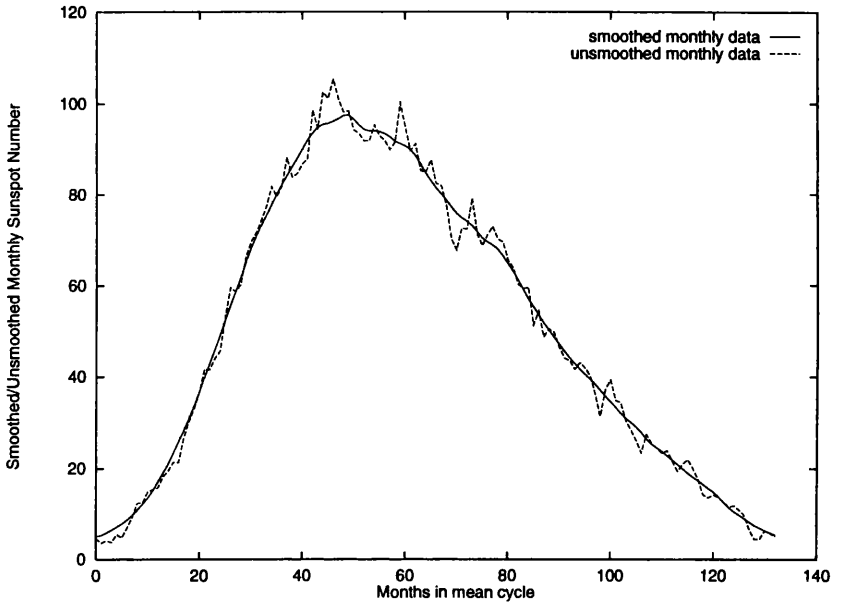


Figure 1.6: The mean cycles obtained in the case of the smoothed and unsmoothed monthly sunspot number.

several prediction accuracies quoted in Kerridge et al. (1989) using this method, produced results which were in very close agreement. Hence it was not thought necessary to test the code with different interpolation schemes. The form of Equation 1.7 is such that the McNish and Lincoln method obviously produces only the next value in the time series. For predictions further ahead than this, the method has to be iterated using its own predictions as subsequent input data. As will be shown later, this is different from the neural network approach where both iterated and non-iterated predictions of more than one timestep ahead are possible.

From this point onwards, subsequent references to the McNish & Lincoln method (or ML method as often abbreviated) should be taken to mean this particular variant of the original paper, unless otherwise stated. To complete this section on the ML method, two examples of typical mean cycles generated by this method are shown. Using smoothed and unsmoothed monthly sunspot numbers as the index under study, the graphs in Figure 1.6 show the mean cycles calculated for each of these datasets. Both follow predominantly the same shape although the unsmoothed data naturally shows up additional variations to the smoothed mean cycle. The maxima of these cycles are around 100, a balance between very small amplitude cycles such as 5 and 6 and the recent higher levels of activity witnessed in cycles 19, 21 and 22.

1.7 Conclusions

The purpose of this chapter was well explained by its title, seeking to give a basic introduction to the nature of solar activity. This is obviously a big picture and so it has not been possible to detail many of the current research areas in solar activity, discussing only those of direct interest. As a result, a great deal of attention has been focussed on describing the International Sunspot Number. As shown, this represents the longest recorded index of solar activity and for this reason alone would stand as the major focus for the predictive methodology of artificial neural networks. The nature of the sunspot cycle, the historical perspective of the first observations and the use of proxy indices to confirm or extend the available record of sunspot data was explained in detail. Solar irradiance variations, in particular the solar 10.7cm flux, have also become important as measures of solar activity. These variations have a much shorter recorded period since some of them require satellite measurements made high in the Earth's atmosphere or beyond.

Once the complicated nature of solar activity variations was discussed, their importance to the near-Earth environment was explained. Of course, from a solar physics point of view, the active Sun is an important area of research in itself but the effect of its activity variations on the terrestrial environment brings the problem closer to home. In this respect specific examples such as solar particle events, geomagnetic storms and the implications for satellite drag were used as case studies of the consequences of solar activity variations.

Finally this chapter provided a summary of the prediction capability of current methods of analysing the sunspot and solar flux cycles. This in itself is quite an extensive subject and so the main thrust of the section was to describe the different types of method which are currently in use, namely statistical and precursor methods. Furthermore the long established statistical method of McNish and Lincoln was explained in detail as it was chosen as the ideal benchmark test for the neural network predictions which will be produced later in this thesis. Before these results can be obtained, it is necessary to introduce the theory of neural network computation and this is carried out in Chapter 2, to which we now proceed.

Chapter 2

Concepts of Neural Network Computation

2.1 Introduction

The aim of this chapter is to attempt to give the reader a basic introduction to the concept of "neural networks" and explain the recent explosion of interest in the range of techniques and applications which come under this general banner headline. While it is obviously beyond the scope of a thesis to present a thorough explanation of the whole field as it currently stands, the author has made it his specific intention to provide the rudimentary ideas which lie at the foundations of the subject. In doing this, it is hoped to explode some of the myths which have sprung up regarding the virtues of neural network models. These exaggerations, which have surely been the result of a struggle for fame and funding, have led to a popular rumour that neural networks can do anything which a traditional computer can, with the added advantage of not needing to be programmed. This idea must be taken in context when consideration is made of the great deal of thought which must go into the design of the network architecture and training strategy for a specific problem. Unfortunately, there are few rules and little experience to help in this area.

2.2 Inspiration

While the term neural networks has been used exclusively in the text to date to describe this field of research, many other expressions are freely used within the literature to refer to the same field. A non-exhaustive list would include neural networks, neural computation, associative networks, collective or network computation and connectionism. Most of the reason for this lies in the twin attitudes to neural network research which have prevailed throughout the 50 years of the history

of the subject. The two roots of the subject can be stated as:

(i) the desire to understand the principles on which the human brain works and to construct a machine that would function in the same way, and

(ii) not to consider the actual plausibility of network systems as brain models but to use them for the analysis of complex systems which are not accessible to sequentially operating programmable computers.

Most of the recent surge in interest in neural networks and neural computation has been devoted to the latter aspect of the subject. It was from neuroscience, however, that the real inspiration came. The theory and application of neural computation have developed from the early research carried out by neurophysiologists into understanding the organisational principles of the mammalian brain. On a microscopic level, due to the organic nature of its constituent parts, the brain is inherently slow when compared with the operational speeds of the central processing units (CPUs) of the most powerful contemporary serial computers. Despite this apparent disadvantage, everyday observations show that even the more modest brains of lower animals will easily perform tasks which are far beyond the range of the most powerful electronic computers. For example, consider that any mosquito can fly around at great speed in unknown terrritory without bumping into objects which may block its path. And a frog's tongue can catch those insects in full flight in a split second. Such achievements require sufficient computational power to recognise complex optical or acoustical patterns in an instant and are beyond the capabilities of present-day computers.
The following properties summarise the many features of the brain which are missing from or would be desirable in artificial systems.

(i) It is robust and fault tolerant with a high degree of error resistivity. This means that the performance is not noticeably degraded by the failure of a single neuron or synaptic connection. This compares very favourably with the operation of a normal computer which will often fail completely in its execution if a single bit of stored information or a single program statement is incorrect.

(ii) It is flexible. It responds to a new stimulus by learning - it does not need to be re-programmed using a programming language such as C, Fortran or Pascal.

(iii) It can deal efficiently with fuzzy, probabilistic, inconsistent or noisy data.

(iv) It is highly parallel, but still small, compact and dissipates little power.

The one area in which it cannot compete with the digital computer is in the execution of simple arithmetic. Perhaps these reasons provide the best motivation for the study of the field of neural networks and neural computation. It offers an alternative computational paradigm to the standard

theory proposed by von Neumann which has been the basis of machine computation to date. The inspiration does indeed come from neuroscience but models and systems are no longer constrained by the requirement of biological realism. It is certainly to be hoped, in the light of all the recent interest in neural networks, and with the parallel advances in computer technology, that neurological research should benefit by the current state of affairs.

2.3 What is a Neural Network?

We start by giving the simplest mathematical description of a neural network. It can be defined as a directed graph with the following properties:

(i) A state variable n_i is associated with each node i.

(ii) A real-valued weight w_{ik} is associated with each connection between two nodes i and k.

(iii) A real-valued bias θ_i is associated with each node i.

(iv) A transfer function $g_i[n_k, w_{ik}, \theta_i, (k \neq i)]$ is defined, for each node i, which determines the state of the node as a function of its bias, of the weights of its incoming connections and of the states of the nodes linked to it by these connections.

2.4 Historical Perspective

The inception of the subject is clearly observed in the work of Warren McCulloch and Walter Pitts in 1943 (McCulloch and Pitts, 1943). While the philosophical debate over human thought processes had been raging since Aristotle's time, this work attempted to give a mathematical description of human nervous activity. They proposed a general theory of information processing based on networks of binary decision elements. These decision elements drew inspiration from the biological neuron although they were much simpler. Specifically, the 'model neuron' takes a weighted sum of its inputs from other units, and outputs a one or a zero according to whether this sum is above or below the threshold value. Using the notation introduced above, this formulation is expressed by:

$$n_i(t+1) = \Theta\left(\sum_j w_{ij} n_j(t) - \theta_i\right) \tag{2.1}$$

where n_i takes the value 1 or 0 and represents the neuron *firing* or *not firing* respectively. The time t is taken to be discrete to simulate the finite regenerative period of real neurons with changes in the state of individual elements of the network taking place at time $t = 0, 1, 2, \ldots$. One unit of time elapses per processing step. Θ is the unit step function given by

$$\Theta(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{2.2}$$

The weight w_{ij} represents the strength of the synapse/connection between neurons j and i and can be positive or negative, corresponding to a excitatory or inhibitory synapse respectively. The threshold θ_i is unit-specific and the weighted sum of inputs must reach or exceed θ_i for the neuron to fire. Though such a network of neurons is apparently simple, McCulloch and Pitts showed that it can, in principle, carry out any computation that an ordinary, programmable, digital computer can perform, although not necessarily as quickly.

In some respects, this first '*neural network*' could be said to have a 'programme code' consisting of the single statement 2.1. This extreme reduction in complexity of a programme obviously has to be compensated for and this is achieved by the substitution of a vast number of processing elements (eg. 10^{11} in the human brain).

Hertz et al. (1991a) discusses the differences which exist between the McCulloch-Pitts formulation and the established behaviour of biological neurons. These include

(i) Real neurons are often not even approximately threshold devices but instead respond to stimulation in a continuous manner. This is referred to as graded response. The essential inclusion in the working model is the nonlinear relationship between the input and output of a cell, as this appears to be a feature common to all cells.

(ii) Many real cells are also thought to perform a nonlinear summation of their inputs, with the possibility also of logical processing (eg AND, OR) within the dendritic tree.

(iii) When it fires, a real neuron produces a sequence of pulses and not a continuous output level. Associating the output with the number in n_i ignores the possibility that the phase of a pulsed sequence, for example, may carry additional information. The phase is not actually thought to be a significant feature in neuronal circuits but expert opinion is not unanimous.

(iv) Neurons do not all have the same fixed delay ($t \rightarrow t+1$), nor are they updated synchronously by a central clock.

(v) The amount of transmitter substance released at a synapse can vary unpredictably.

By a simple generalisation of the McCulloch-Pitts formulation, features (i) and (iv) are easily included in the model. We write

$$n_i := g(\sum_j w_{ij} n_j - \theta_i) \tag{2.3}$$

The output of the unit is now continuous-valued and n_i is called the **state** or **activation value** of the unit. The unit step function Θ is replaced by a general function g(x). This is called the **activation function** or **transfer function**. Equation 2.3 now provides an updating rule for each individual unit in the network but the explicit time dependence has been removed. The updating can now take place asynchronously, ie in random order at random times. The notation $:=$ has been

borrowed from Hertz et al. (1991a) to indicate that the left hand side of the equation is assigned the value on the right after updating but that the equality is not true continuously. Some research has also been concerned with simulating feature (v) through the theory of stochastic neurons.

Before continuing with this brief history, it is worth mentioning briefly the two processes of brain function which it is most desired to understand. One is the process of memory while the other concerns the brain's ability to learn from examples. These issues must in reality be related as the method of memory recall must surely be linked to how the information was learned and stored initially. It is clear from the history, however, that most work has been concerned with modelling one or other process.

In the late 1950's and early 1960's, attention was focussed on the problem of how to find appropriate connection strengths w_{ij} which solve a particular cognitive task. Here, following the definition in Müller and Reinhardt (1991), 'cognitive task' is used in a generalised sense to mean any task requiring digital, or even analogue, information processing, such as the recognition of specific optical or acoustical patterns. Caianiello (1961) proposed a "*learning*" algorithm that would determine the synaptic strengths of a neural network. Rosenblatt (1958) introduced a specific type of neural network called the **perceptron**. This was to stimulate the first real interest in neural networks as a major research field. Further work by Rosenblatt (1962) and Block (1962) culminated in the Perceptron Convergence Theorem. This theorem showed that, assuming there was a solution to a problem, that the perceptron would learn it within a finite number of steps. The work of this time propelled neural network techniques, principally in the form of the perceptron, into their first "golden age". At this time, a hundred algorithms were proposed as it seemed that neural networks could do anything. The ride lasted until 1969 with the publication of 'Perceptrons' by Minsky and Papert (1969). Suddenly, from the great heights which had been reached, research in this major area of neural network studies became very unfashionable and consequently, unfunded in the main. It would not be fair to place the blame entirely on this one publication as doubts over the true abilities of the perceptron had been coming to the fore in the months and years immediately preceding this publication. 'Perceptrons', indeed, is a major work as shown by the initial review of Alan Newell in Science, who began "This is a great book". The particular point about Minsky and Papert's work was that while the earlier papers by Rosenblatt and others dealt with the possibilities of neural networks with little attention to their limitations, Minsky and Papert produced a manuscript which placed the computational limitations of perceptrons on a strict foundation. In particular, they produced examples of very simple problems, most notably the logical Exclusive-Or function, which the perceptron could not solve. Perhaps the most crucial statement regarding its impact on future work came in the conclusions where Minsky and Papert conjectured that the limitations of the simple perceptron would also hold for variants of this model, most notably the extension to multilayer systems. Such systems had also been

proposed by Rosenblatt in his earlier work (Rosenblatt, 1962). While multilayer systems could in principal solve the XOR among other problems, it was thought hard to envisage a learning algorithm applicable to these more complicated systems. It now appears that this conjecture was false, with the independent (re)-discovery by Parker (1985), Le Cun (1985) and Rumelhart, McClelland et al (1986) (see Rumelhart et al., 1986a, Rumelhart et al., 1986b and Rumelhart et al., 1986c) of a learning algorithm which has become known as the "back-propagation of errors" algorithm. This algorithm appears to have been proposed first in the doctoral thesis of Werbos (1974) and is a generalisation of the related Widrow-Hoff rule (Widrow and Hoff, 1960). The recent resurgence in the fortunes of neural networks is largely a result of the back-propagation algorithm with most research concentrated around its application to various types of problems.

Not all research was inactive in the intervening period between the fall and rise again of perceptron models. A major theme that received attention throughout the 1970's was **associative content-addressable memory**. As previously mentioned, this followed from the desire to model distributed memory within the brain. While standard computers operate a wholly address-oriented system of information storage where every bit of information is assigned an address in memory from where it can be recalled, the brain itself obviously does not behave as such. For example, nobody recalls the 5632nd person whose name they heard. To this extent memory is surely associative so that recall of a name, to continue the example, is made by association with a face or occupation etc. This is also referred to as content-addressable as it is from knowing some aspect of the content of the memory which is sufficient to trigger recall of the information required.

Neural network models of associative memory have been studied since Taylor (1956). In the 1970's, this work was expanded and continued by Anderson (1968, 1970, Anderson and Mozer 1981), Marr (1969,70,71) and Kohonen (1974-89) among others. In particular, Stephen Grossberg (1967-87) has been exceptionally productive in this field and has produced his own general reformulation of memory and learning in neural theory.

The other main line of development was initiated by Little (1974) and continued by Little and Shaw (1978) and Hopfield (1982). This work considered the similarity between the original McCulloch-Pitts formulation of a neuron and systems of elementary magnetic moments or "spins". In these systems, the spin s_i at each lattice site i can take only one of two different orientations, up or down, denoted by $s_i = +1$ and $s_i = -1$ respectively. The analogy is obvious when one associates each spin with the state of a neuron n_i as either 1 or 0. This work became especially fruitful due to extensive statistical mechanics studies in the understanding of the thermodynamic properties of disordered systems of spins. No further attempt will be made to cover this aspect of the theory. The reader is referred to the text by Müller and Reinhardt (1991) for a full physical description, most notably in Section II of that book where the statistical mechanics description of neural networks is discussed in depth.

Indeed most attention within this thesis will subsequently be concerned only with the theory and applications of the learning algorithms, most notably "back-propagation of errors", of perceptrons and multilayered variants of this model. It is the applications of this type of neural network system we will be concerned with studying. For the sake of completeness, a brief discussion of associative content-addressable memory will be made shortly, since in many respects this provides a good example of the simplest application of neural networks. For more detailed descriptions, the texts of Hertz et al. (1991a) and Müller and Reinhardt (1991), and references contained therein are particularly recommended as they offered the present author the greatest part of his current understanding of neural network techniques. Anderson and Rosenfeld (1988) have collected many of the historically significant papers in one volume together with their own comments placing the work in context with parallel developments. This volume was drawn upon heavily for this section.

2.5 Main Issues behind Parallel Processing

At this point enough introduction has been covered to enable some of the major questions regarding the performance of neural networks to be posed. It is these questions which we will be dealing with continually in the later chapters with the application of real data problems to a neural network model.

- What is the best architecture? Should the network be divided into layers and if so how many are required? How many units should different layers contain and how should they be connected? Do different activation functions play a role and what sort of updating should be employed, synchronous or asynchronous?

- How can a network be programmed? Can it learn a task and if so with how much information does it need to be presented to give a good performance? How long does it take to learn given that multiple presentations of the information may be required? What is the best method of learning and can it learn in real time while functioning or should the training phase be carried out independently?

- Finally, having learned a task, how robust is it to missing or noisy information? Can it 'generalise' outwith its learnt examples to carry out related but unknown tasks?

Other questions to be raised concern the hardware implementation of neural networks and how they compare to the simulations made by computer software.
The mention of generalisation explains some of the great excitement that surrounds neural networks. In some cases they have been found to be able to interpolate and extrapolate to form a complete relationship, having been given only examples of the relationship. How the network

generalises, to what problems it can produce these generalisations and whether this generalisation is 'sensible' or not remains a major question in research.

2.6 Programming or Learning

The solution to all problems analysed by a neural network exists in the choice of the connection strengths or weights w_{ij}. In some cases, this choice can be made a priori to predesign information in the network. Most times, however, it is hoped to "teach" or train the network to perform a particular cognitive task by iterative adjustments of the w_{ij}'s. The two main ways of carrying this out are:

(i) **Supervised learning** where training takes place through direct comparison of the network's output with answers which are known for the examples in the training set. From an initially random network set-up, the network is fed an input pattern from which it computes an output pattern. This output is then "corrected" through presentation of the desired output which matches with the input pattern. Through many input-output pairs and repeated presentations, the network minimises the differences between its computed values and the correct outputs by appropriate changes in the weight connections. Then the network is said to have 'learnt' the problem, once the training set is known to high accuracy. At this point it is of great interest to present the network with input patterns outwith the training set and test its ability to generalise what it has learnt.

(ii) **Unsupervised learning** where the only available information lies in the correlations of the input data. There is no right or wrong correction process and the network must discover for itself interesting features or correlations in the input data. Intuitively, it can be seen that this requires a certain **redundancy** in the input data, otherwise any patterns or features which showed up would necessarily be assumed to be random noise. This type of learning allows for many interesting applications and may yet prove fruitful in an astronomical context for recognising, for example, galaxy clustering.

In essence, this is the very hope of applying these techniques to complex problems. Instead of having to specify every detail of a calculation, particularly in cases where these details are not known exactly, perhaps there is scope for an approach based on learning by example.

2.7 Basics of Associative Memory – the Hopfield Model

Hertz et al. (1991a) defines an associative memory as follows:

Store a set of p patterns ξ_i^μ, each containing N bits of information ($i = 1, \ldots, N; \mu = 1, \ldots, p$), such that on presentation of a new pattern n_i, the network produces as output the pattern ξ_i^μ which resembles most closely n_i.

This task can easily be carried out conventionally by computing the Hamming distance between the presented pattern and those stored. (The Hamming distance between two binary numbers is just the number of bits in which the numbers differ). For large numbers and many patterns, this proves computationally expensive. The aim of associative memory networks is to get a McCulloch-Pitts network to solve the problem. So if the initial configuration is $n_i = \zeta_i$ then can a set of w_{ij}'s be chosen which will make the network go to the state with $n_i = \xi_i^{\mu_o}$, where pattern number μ_0 is the smallest distance from ζ_i. The strength of this formulation lies in the fact that the memory can be insensitive to small errors, for example "$E = mc^3$" will recall Einstein despite the input error, and furthermore, the network will never retrieve a linear combination of two responses but will choose the nearest fit. Particular applications of such a memory are in image reconstruction or retrieval of bibliographic information from partial references.

The formulation of this problem is usually written using the McCulloch-Pitts equations (2.1 and 2.2) with -1 and 1 as the activation values instead of 0/1. The sign function is used for this (see Equation 2.13) and we consider only the case of asynchronous updating where the units are updated one at a time according to the equation

$$n_i := sgn(\sum_j w_{ij} n_j) \tag{2.4}$$

Thresholds are not included in this model. The choice of connection weights is considered first in the case of having only one pattern to be memorised. The condition for this pattern to be stable is

$$sgn(\sum_j w_{ij} n_j) = n_i \qquad \text{for all i} \tag{2.5}$$

because then 2.4 makes no changes. This is true if

$$w_{ij} \propto n_i n_j \tag{2.6}$$

since $n_j^2 = 1$. The constant of proportionality is taken as $1/N$ where N is the number of units in the network.

Importantly, it is also the case that if a number of bits (less than half) of the starting pattern are wrong, they will be overwhelmed in the sum for net input

$$h_i = \sum_j w_{ij} n_j \tag{2.7}$$

by the majority which are right, and $sgn(h_j)$ will still give the right output. Thus the network will correct errors as an initial configuration near to n_i will relax to n_i fairly quickly. This can be

extended to the case of recalling many patterns by making w_{ij} a superposition of terms like 2.6, one for each pattern

$$w_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} n_i^\mu n_j^\mu \tag{2.8}$$

where p is the total number of stored patterns.

This rule is usually called the "Hebb Rule" due to the similarity with a hypothesis made earlier by Hebb (1949). An associative memory model using the Hebb rule for all possible pairs i, j using binary units and asynchronous updating is generally called the Hopfield Model after Hopfield brought all the pieces together in his infuential paper (Hopfield, 1982). This section has only involved a cursory explanation of this model and drew heavily on the review by (Hertz et al., 1991a). Other reviews were carried out by Cowan and Sharp (1988b) and Cowan and Sharp (1988a).

2.8 Perceptrons and Learning Algorithms

Whereas in some simple tasks involving small networks (and hence small numbers of connections) and few patterns, it may be possible to choose a priori the weight connections w_{ij} which solve a particular problem, this is not practical for most situations. In these cases, it is hoped to solve the problem by choosing initial random weights and iterating progressively, by means of a suitable algorithm, until a solution is found.

The two ways in which a network can learn, either (i) **supervised learning** or (ii) **unsupervised learning**, were described in more detail in Section 2.6. From this point, the discussion will be limited to supervised learning within **layered feed-forward networks**. Such an architecture describes a network with layers of units, arranged so that every unit is only connected to a unit in a layer above it, ie no backward or sideways connections are allowed. This is usually restricted further to allow only connections between units in one layer and the layer immediately after it. There is another class of networks, still employing supervised learning but not strictly feed-forward, which are referred to as **recurrent networks**. The applications of these architectures will be discussed briefly later (see Section 7.3.2).

When Rosenblatt proposed the perceptron in 1962 (Rosenblatt, 1962), it was based on the layered feed-forward architecture. It consisted of a layer of input units, an output layer where the results of the computation are given and in between intermediate layers of units referred to as **hidden** layers or **hidden** units since there are no outside connections. In the simplest case there is no hidden layer and this architecture was called the simple perceptron by Rosenblatt. Since the input layer carries out no computation but merely presents the input data to the network, it is frequently ignored when counting the number of layers in a network. Subsequently the simple perceptron is equivalent to a 1-layer network while a network with one hidden layer of units is referred to as

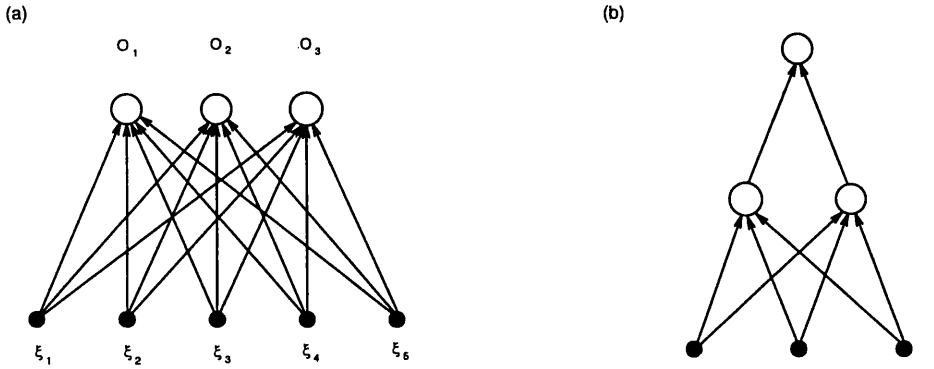


Figure 2.1: Perceptrons. (a) A simple perceptron which (by definition) has only one layer. (b) A two-layer perceptron. Inputs, shown as solid circles, perform no computation and are not included in the count of layers.

a 2-layer network. Figure 2.1 shows examples of each. This notation will be followed from now on although it should be stated that while the majority of authors use this method of counting layers, it is by no means entirely standard. Note then that an $N-$layer perceptron has N layers of connections and $N-1$ intermediate or hidden layers.

Consider first the case of a 1-layer network, that is a layer of input units connected directly to the output layer. In all subsequent discussions ξ_k will denote the value in the k^{th} unit of the input layer ($k = 1, \ldots, N_{in}$), O_i denotes the output of the i^{th} unit in the output layer ($i = 1, \ldots, N_{out}$) and N_{in} and N_{out} represent the number of units in the input and output layers respectively. This convention of using the subscripts k and i for units in the input and output layers respectively will be retained throughout. The superscript μ will be used in addition to identify the different input-output patterns contained in the training set ($\mu = 1, \ldots, p$) , where p is the total number of training patterns. For every input pattern ξ_k^μ in the training set, there is a corresponding known output pattern T_i^μ. The network computes an output given by

$$O_i^\mu = g(h_i^\mu) = g(\sum_{k=1}^{N_{in}} w_{ik}\xi_k) \tag{2.9}$$

Depending on the form of the activation function $g(h)$ as a threshold or continuous function, this compares with either 2.1 or 2.3 of the McCulloch-Pitts formulation respectively. Bias values can easily be introduced as before giving

$$O_i^\mu = g(h_i^\mu - \theta_i) = g(\sum_{k=1}^{N_{in}} w_{ik}\xi_k - \theta_i) \tag{2.10}$$

If the activation function is a threshold function as in the original McCulloch-Pitts formulation 2.1, then the θ's are usually referred to as threshold values.

From Equations 2.9 and 2.10, it is clear that the output is an explicit function of the inputs ξ_k and this is a general feature of all layered feed-forward network architectures. The desired result

of this computation is that

$$O_i^\mu = T_i^\mu \tag{2.11}$$

for all patterns μ. Working from Equation 2.9 it is clear that this can only be achieved if there exists a suitable choice of weights w_{ik} to satisfy 2.11.

The auto-association problem of storing, in memory so to speak, the patterns ξ_k^μ so that presentation of a pattern produces an identical output is contained within this formalism. It is however the **hetero-association** problem which is of most interest and which is usually covered by this approach. In this case N_{out} does not need to equal N_{in} and if they do then

$$O_i^\mu \neq T_i^\mu \tag{2.12}$$

It can be shown that for a simple perceptron to be able to solve a problem then the problem itself must be **linear separable**, in the case of threshold units and **linearly independent** when linear units are used. The case involving nonlinear activation functions will be mentioned later. A linear separable problem is one in which a plane can be found in ξ space which divides the inputs which have positive or negative outputs. This can be demonstrated as follows. From Equation 2.9, consider the sign function $\mathrm{sgn}(x)$ as the activation function, where

$$\mathrm{sgn}(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ -1, & \text{otherwise} \end{cases} \tag{2.13}$$

and assume the outputs take the values $T_i^\mu = \pm 1$. Since the output units are independent, the subscript i can be dropped allowing the condition 2.11 to be re-written (using 2.9 and 2.13) as

$$\mathrm{sgn}(\mathbf{w}.\xi^\mu) = T^\mu \tag{2.14}$$

where $\mathbf{w}.\xi^\mu$ is the scalar product of the two vectors $\mathbf{w} = (w_1, w_2, \ldots, w_{N_{in}})$ and $\xi^\mu = (\xi^1, \ldots, \xi^\mu_{N_{in}})$. Condition 2.14 requires that $\mathbf{w}$ be chosen such that the projection of pattern ξ^μ onto it has the same sign as T^μ. But the boundary between positive and negative projections onto $\mathbf{w}$ is the plane $\mathbf{w}.\xi^\mu$ through the origin perpendicular to $\mathbf{w}$. This shows the requirement for the problem to be linearly separable. (Reintroducing the threshold values θ means that the plane need not pass through the origin). When there is more than one output unit, a plane must be found for each one.

In the case where the activation function is taken as $g(h) = h$, so that it is continuous and differentiable, this describes a situation described by linear units. Analysis of this type of network (see eg, Hertz et al., 1991a) gives the sufficient condition that the input patterns be linearly independent to ensure solvability of a problem. This condition is distinct from linear separability for threshold units as linear independence implies linear separability but the reverse is not true. Many of the Boolean logical problems studied using threshold networks do not satisfy the linear independence condition, with typically p (No. of patterns) $> N_{in}$ (No. of input units), but

ξ_1	ξ_2	O
0	0	0
1	0	1
0	1	1
1	1	1

(a)

ξ_1	ξ_2	O
0	0	0
1	0	1
0	1	1
1	1	0

(b)

Table 2.1: The truth tables for (a) the OR function and (b) the XOR function.

are nevertheless linear separable. Nonlinear units are those which employ a nonlinear activation function $g(h)$, although it is still a differentiable function. The conditions for the existence of a solution are the same as for linear units namely the linear independence of the input patterns.

2.8.1 Examples of Simple Threshold Perceptrons

At this point it is constructive to illustrate the previous description with a couple of examples, one of a problem which is solvable and one which isn't.

The Boolean OR Function

This is a function of two 0/1 variables so a perceptron with two input units ξ_1 and ξ_2 is used. Using the unit step function (see Equation 2.2) which outputs 0/1, the function gives a 1 if either (or both) input are on (ie have value 1) and a 0 only if both units are off. The truth table for the function is given in Table 2.1(a).

In (ξ_1, ξ_2) space, the function is shown in Figure 2.2(a). The OR function is thus linearly separable with the dividing plane (dashed line) perpendicular to the weight vector $\mathbf{w}$.

The Exclusive–OR (XOR) Function

This function differs from the OR function in that it gives a 1 output if ξ_1 or ξ_2 are on but not both. The full truth table is shown in Table 2.1(b) and Figure 2.2(b) the corresponding diagram. It is obvious from this diagram that no plane (line) can separate the two types of point (• and ○). The XOR problem is the simplest form of the N-input parity function which was discussed by Minsky and Papert (1969) and shown to be insoluble by a simple perceptron. This can also be shown algebraically by writing out explicitly the Equations 2.10 for each of the 4 input patterns. Dropping the output subscript i, these are

$$w_1 + w_2 < \theta \tag{2.15}$$

$$w_2 > \theta \tag{2.16}$$

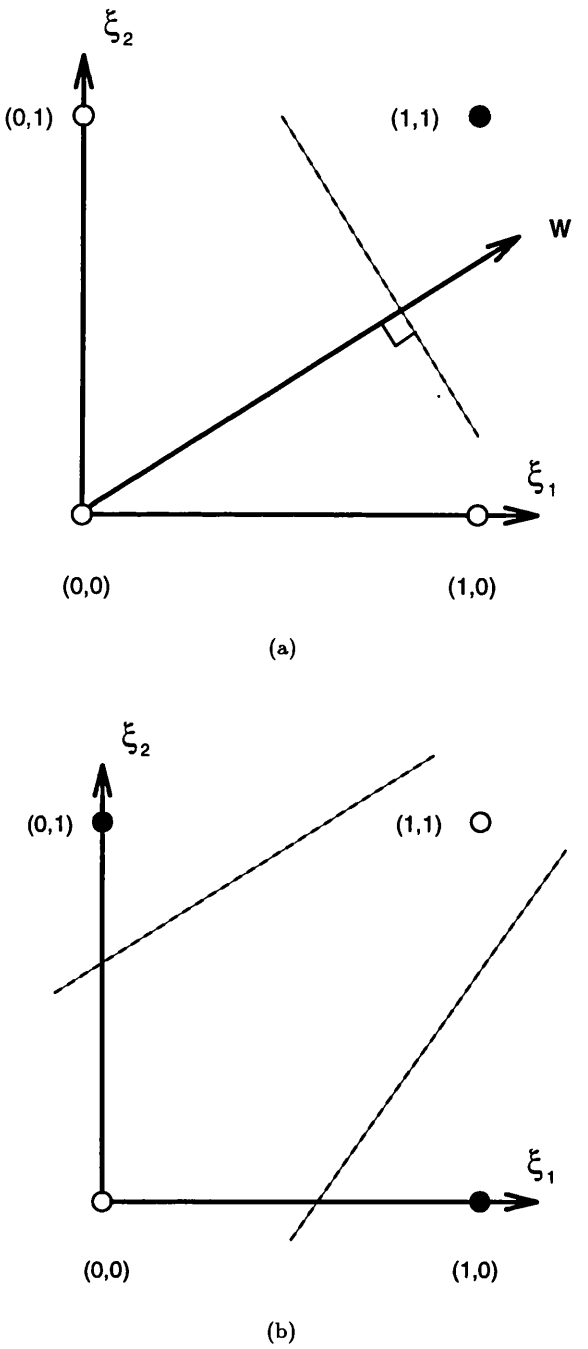


Figure 2.2: The input-output pair for (a) the OR function and (b) the XOR function when plotted in (ξ_1, ξ_2) space.

$$w_1 > \theta \tag{2.17}$$

$$\theta > 0 \tag{2.18}$$

From the first two of these four equations we require $w_1 < 0$ while the last two give $w_1 > 0$ as the necessary condition. Since this is obviously incompatible, it confirms that the XOR is insoluble.

2.8.2 Learning Algorithms

Rosenblatt (1962) and Block (1962) proposed a simple learning algorithm for their threshold networks based on the premise of asking whether or not the computed output matches the target, and if not, the weight connections are supplemented by a factor proportional to the difference between the two. An important feature of their work was the proof of the Perceptron Convergence Theorem (see for example , Rosenblatt, 1962, Block, 1962, Minsky and Papert, 1969). This theorem showed that, provided a solution exists, a perceptron with a specific learning algorithm would find a solution within a finite number of steps.

At around the same time, Widrow and Hoff (1960) proposed a learning algorithm for simple networks with linear units. In essence an **error** or **cost function E** is defined as

$$E[\mathbf{w}] = \frac{1}{2}\sum_{i\mu}(T_i^\mu - O_i^\mu)^2 = \frac{1}{2}\sum_{i\mu}(T_i^\mu - \sum_k w_{ik}\xi_k^\mu)^2 \tag{2.19}$$

which depends only on the weights and the input patterns and which goes to zero when condition 2.11 is satisfied. A search must be carried out in **w** space to minimise the error and Widrow and Hoff proposed the method of **gradient descent** as the most useful method to achieve this. This changes each weight w_{ik} by an amount Δw_{ik} proportional to the gradient of **E** at the current location.

$$\Delta w_{ik} = -\varepsilon\left(\frac{\partial E}{\partial w_{ik}}\right) = \varepsilon\sum_\mu (T_i^\mu - O_i^\mu)\,\xi_k^\mu \tag{2.20}$$

where we introduce ε as the constant of proportionality known as the *learning rate*. So for each individual input pattern we can define the weight changes by

$$\Delta w_{ik} = \varepsilon\delta_i^\mu \xi_k^\mu \tag{2.21}$$

where the errors (or deltas) are $\delta_i^\mu = (T_i^\mu - O_i^\mu)$. This learning rule is called the **delta rule** or **Widrow-Hoff rule**.

It can easily be generalised to cover nonlinear units ie those with a nonlinear (but differentiable) activation function g(h). A sigmoid function is the usual choice for g(h) with either the logistic function or the hyperbolic tangent being utilised depending on the output range desired. The former scales its output to between 0 and 1 while the tanh(h) gives a $-1/1$ output range. The reason for this type of function is discussed in more detail in Section 2.9.2. Equations 2.19 and

2.20 become respectively,

$$E[\mathbf{w}] = \frac{1}{2}\sum_{i\mu}\left(T_i^\mu - g(\sum_k w_{ik}\xi_k^\mu)\right)^2 \tag{2.22}$$

$$\Delta w_{ik} = -\varepsilon\left(\frac{\partial E}{\partial w_{ik}}\right) = \varepsilon\sum_\mu (T_i^\mu - O_i^\mu)\, g'(h_i^\mu)\xi_k^\mu \tag{2.23}$$

$$\text{where } h_i^\mu = \sum_k w_{ik}\,\xi_k^\mu \tag{2.24}$$

So again for each individual pattern we define the weight changes by

$$\Delta w_{ik} = \varepsilon\delta_i^\mu\xi_k^\mu \tag{2.25}$$

where this time $\delta_i^\mu = (T_i^\mu - O_i^\mu)\, g'(h_i^\mu)$. In 1-layer networks, nonlinear units represent no real advantage over a linear response, with the network still limited to the condition of linear independence in the input patterns. This is obvious since the nonlinear case is equivalent to the linear one except with the target values replaced by $g^{-1}(T_i^\mu)$. (Since the activation function is sigmoidal, it is monotonic and hence invertible, except possibly at the endpoints of the range.)

In multi-layer networks, the use of nonlinear units becomes much more useful. A multi-layer linear network is equivalent to a 1-layer network in the computation it performs (since a linear transformation of a linear transformation is a linear transformation). Multi-layer nonlinear networks do not suffer from this restriction. Indeed, it was known in the 1960's that a 3-layer network could provide a solution to the XOR problem among others. The lack of existence of a learning algorithm was a severe limitation in their application at that time. Minsky and Papert (1969) conjectured that the limitations of simple perceptrons would also hold true for multi-layer systems. This, in part, led to the decline in interest in this research until the 1980's when three independent groups (re)-discovered a learning algorithm for multi-layer networks and the revival was instigated.

2.9 Multi-layer Feed-forward Networks

The treatment here is still restricted to feed-forward networks. We will also consider only 2-layer networks, that is networks with only one hidden layer, although it will be obvious that these results may be extended quite easily for larger architectures. The earlier notation of ξ_k/O_i for input/output network values is retained and V_j will denote the values of the N_{hid} hidden units ($j = 1, 2, \ldots, N_{hid}$). The subscripts k, j and i are used as standards for units in the input, hidden and output layers respectively. The weight connections between the input and hidden layers are denoted by w_{jk} and between the hidden and output by W_{ij}. If necessary, bias values θ_j are included on the hidden units. They may also be used on the output units but this is unnecessary. The network calculations proceed as follows, given input pattern ξ_k^μ. Hidden unit j receives as input

the quantity

$$h_j^\mu = \sum_k w_{jk}\xi_k^\mu \tag{2.26}$$

It calculates its output value V_j^μ based on the calculation

$$V_j^\mu = g(h_j^\mu - \theta_j) = g(\sum_k w_{jk}\xi_k^\mu - \theta_j) \tag{2.27}$$

This value is passed through the W_{ij} weight connections to the subsequent layer, in this discussion the final output layer. Similar calculations to before yield

$$h_i^\mu = \sum_j W_{ij}V_j^\mu \tag{2.28}$$

as input to the final layer and calculates the final output as

$$O_i^\mu = g(h_i^\mu) = g(\sum_j W_{ij}V_j^\mu) \tag{2.29}$$

In this type of network the Widrow-Hoff style learning rule cannot be directly implemented since, although the error function can be formed as before, many connection weights give rise to this error, not just the ones at the output layers. The problem exists in calculating the error for the hidden units which are embedded inside the network. The work of the Parallel Distributed Processing Group at the University of California, San Diego becomes important at this stage. It has already been acknowledged that the 'back-propagation of errors' algorithm was also discovered independently of this Group (see Werbos, 1974, Parker, 1985, Le Cun, 1985, Bryson and Ho, 1969). The publication of **Parallel Distributed Processing: Explorations in the Microstructures of Cognition, 1986** (two volumes), however, of which D.E. Rumelhart and J.L. McClelland, two prominent members of the PDP Research Group in the early and mid 1980's, were editors, is probably still the most detailed single-source of information on the field of neural networks.

In Rumelhart et al. (1986a), the authors restate the work of Minsky and Papert (1969) who covered the capabilities and limitations of 1-layer and multi-layer systems. The lack of a rule for learning in networks with hidden units which has similar power as the Perceptron Convergence Procedure or the delta-rule variation due to Widrow and Hoff is particularly noted. Rumelhart et al. (1986a) state three possible responses to this lack. One response is represented by competitive learning in which simple *unsupervised* learning rules are employed to develop useful hidden units. This type of approach is promising but there is no external force to *ensure* that hidden units appropriate for the required mapping are developed. The second type of approach is simply to *assume* an internal representation that seems reasonable on some a priori grounds. For more details of this the reader is referred to McClelland and Rumelhart (1981) and Rumelhart and McClelland (1982). Their third approach is the main aim of the work and concerns the "attempt to develop a learning procedure capable of learning an internal representation adequate for performing the task at hand".

The procedure which they developed was a clear generalisation of the Widrow-Hoff (delta) rule and so initially they called it the **generalised delta rule**. It is now more commonly known as the **back-propagation of errors learning algorithm**.

2.9.1 The Back-Propagation of Errors Algorithm

Following the treatment of Rumelhart et al. (1986a), the generalised delta rule is derived for semi-linear activation functions in a network with hidden layers. A semi-linear function is used to mean a nonlinear function in which the output of a unit is a nondecreasing and differentiable function of the net total output

$$h_I = \sum_J w_{IJ} o_J \tag{2.30}$$

Thus a semilinear function is one in which

$$o_J = g(h_J) \tag{2.31}$$

and g is differentiable and nondecreasing. In these equations and the following, the use of capital letters as suffices will indicate that the specific layer significance attached to the letters k, j and i has been relaxed. We consider the same sum-squared error function $E[\mathbf{w}]$ defined earlier, namely

$$E[\mathbf{w}] = \frac{1}{2} \sum_{i\mu} (T_i^\mu - O_i^\mu)^2 \tag{2.32}$$

only in this instance the dependence of the O_i's on the input values ξ_k is given by Equations 2.26 to 2.29. Following gradient descent, we set

$$\Delta w_{IJ} \propto -\frac{\partial E}{\partial w_{IJ}} \tag{2.33}$$

It is useful to see this derivative as the product of two parts: one part reflects the change in error as a function of the change in input to the unit and the other part represents the effect of changing a particular weight on the input to the unit. Hence we write,

$$\frac{\partial E^\mu}{\partial w_{IJ}} = \frac{\partial E^\mu}{\partial h_I^\mu} \frac{\partial h_I^\mu}{\partial w_{IJ}} \tag{2.34}$$

From Equation 2.30 we obtain the second factor as

$$\frac{\partial h_I^\mu}{\partial w_{IJ}} = \frac{\partial}{\partial w_{IJ}} \sum_J w_{IJ} o_J^\mu = o_J^\mu \tag{2.35}$$

We now define

$$\delta_I^\mu = -\frac{\partial E^\mu}{\partial h_I^\mu} \tag{2.36}$$

Thus we rewrite Equation 2.34 as

$$-\frac{\partial E^\mu}{\partial w_{IJ}} = \delta_I^\mu o_J^\mu \tag{2.37}$$

Hence to implement gradient descent in E, the weight changes should be made according to

$$\Delta w_{IJ} = \varepsilon \delta_I^\mu o_J^\mu \tag{2.38}$$

This is as in the standard delta rule although the form of δ_I^μ for each unit u_I in the network must be found explicitly.

To compute $\delta_I^\mu = -\frac{\partial E^\mu}{\partial h_I^\mu}$, apply the chain rule and again write the partial derivative as a product of two factors, one reflecting the change in error as a function of the output of the unit, the other the change in the output as a function of the changes in the input. Thus

$$\delta_I^\mu = -\frac{\partial E^\mu}{\partial h_I^\mu} = -\frac{\partial E^\mu}{\partial o_I^\mu}\frac{\partial o_I^\mu}{\partial h_I^\mu} \tag{2.39}$$

From Equation 2.31,

$$\frac{\partial o_I^\mu}{\partial h_I^\mu} = g_I'(h_I^\mu) \tag{2.40}$$

which is the derivative of the activation function evaluated at the input h_I to that unit. [The subscript I on the g indicates that it is possible to have a system which has a different activation function for each unit. In most, if not all, situations this is not utilised with a single activation function applying to all units beyond the input layer. Henceforth, this particular subscript will be omitted]. To compute the first factor in Equation 2.39, two cases must be distinguished. First assume that unit u_I is an output unit of the network. In this case, from the definition (see Equation 2.32) of $E[\mathbf{w}]$,

$$\frac{\partial E^\mu}{\partial o_I} = -(T_i^\mu - O_i^\mu) \tag{2.41}$$

Substituting back into Equation 2.39, we obtain

$$\delta_I^\mu = (T_i^\mu - O_i^\mu)g'(h_I^\mu) \tag{2.42}$$

for any output unit u_I. This is equivalent to the previous definition of δ_I^μ when $o_I^\mu = h_I^\mu$ due to the linearity of unit u_I.

If u_I is *not* an output unit but one internal to the network, again the chain rule is implemented to write

$$\begin{aligned}\sum_K \frac{\partial E^\mu}{\partial h_K^\mu}\frac{\partial h_K^\mu}{\partial o_I} &= \sum_K \frac{\partial E^\mu}{\partial h_K^\mu}\frac{\partial}{\partial o_I}\sum_J w_{JK}o_J^\mu \\ &= \sum_K \frac{\partial E^\mu}{\partial h_K^\mu}w_{KI} = -\sum_K \delta_K^\mu w_{KI}\end{aligned} \tag{2.43}$$

This time substituting back into Equation 2.39 yields

$$\delta_I^\mu = g'(h_I^\mu)\sum_K \delta_K^\mu w_{KI} \tag{2.44}$$

whenever u_I is not an output unit.

These results can be summarised in three equations, noting that the generalised delta rule has the same form as the standard delta rule of Equation 2.21. This states that the weight of each connection should be changed by an amount proportional to the product of an error, or delta δ, available to the unit receiving input along that connection. That is

$$\Delta w_{IJ}^{\mu} = \varepsilon \delta_I^{\mu} o_J^{\mu} \tag{2.45}$$

This is the first equation. The other two state the error value. Essentially this error is determined recursively starting from the output units. If a unit is an output unit then the error, or delta, is given by

$$\delta_i^{\mu} = (T_i^{\mu} - O_i^{\mu}) g'(h_i^{\mu}) \tag{2.46}$$

Finally the delta value for a hidden target (for which there is no specified target) is determined recursively in terms of the deltas of the units to which it connects directly and the weights of these connections.

$$\delta_J^{\mu} = g'(h_J^{\mu}) \sum_K \delta_K^{\mu} w_{KJ} \tag{2.47}$$

Equation 2.45 is often written in general terms as

$$\Delta w_{pq} = \varepsilon \sum_{\mu} \delta_{\text{output}} \times V_{\text{input}} \tag{2.48}$$

where output and input refer to the two ends p and q of the connection concerned. V stands for the appropriate input-end activation from a hidden unit or a real input. The form of δ takes either of the forms 2.46 or 2.47 depending on whether it refers to the last layer of output connections or for any other layer respectively.

Equation 2.47 allows for the determination of δ for a given hidden unit u_q in terms of the δ's of the units u_p that it feeds. The coefficients are just the usual 'forward' weight connections w_{pq} except here they are propagating the errors backwords (instead of signals forward): hence the name "**back-propagation of errors**" or **back-propagation** for short. Therefore a bidirectional form of the same network is used to compute both the output values and the deltas.

Although the general form of the update rule (Equation 2.48) is often written expressed as a sum over all patterns $P(\mu = 1, \ldots, P)$, it is usually applied *incrementally*: a pattern μ is presented at the input and then all the weight updates are carried out before the next pattern is considered. The alternative **batch mode**, taking 2.48 literally, requires additional local storage for the accumulating weight changes and although the relative effectiveness of the two approaches depend on the problem, Hertz et al. (1991a) note that the incremental approach seems superior in most cases as it decreases the cost function at each step and lets successive steps adapt to the local gradient.

From the observation that $\frac{\partial E}{\partial w_{pq}} = \sum_{\mu} \frac{\partial E^{\mu}}{\partial w_{pq}}$, Rumelhart et al. (1986a) accept that the generalised delta rule implements a gradient descent in E, strictly speaking only in the case of batch mode

updating. For sufficiently small values of the constant of proportionality (the learning rate ε), however, the departure from gradient descent by the incremental method is negligible and the delta rule will approximate gradient descent to find a set of weights minimising the sum-squared error function.

There is no reason why the back-propagation algorithm cannot be implemented in a network where connections exist bridging more than one layer, providing no backward or lateral steps are included. In this case, a unit receives two different types of error, that from direct comparison with the target and that passed from the other output units whose activation it affects. Consequently the weight changes dictated by direct comparison and those which are propagated back are simply added together.

2.9.2 Form of the Activation Function

Throughout the previous discussion, it has been essential to include a nonlinear, differentiable activation function. Taking into account the earlier remarks concerning semilinear functions, it is usual to use the sigmoid logistic function

$$g(h) = f_\beta(h) = \frac{1}{1 + \exp(-2\beta h)} \tag{2.49}$$

for outputs in the range 0/1. For a ± 1 range, an equivalent form is

$$g(h) = \tanh \beta h \tag{2.50}$$

The reasons for this choice reside in the statistical mechanics theory of neural networks. for more details see Müller and Reinhardt (1991), Chapter 4 and (Hertz et al., 1991a), Appendix A.3. The factor β is often set to 0.5 in 2.49 and 1 for version 2.50. These choices of activation function have the added advantage that their derivatives are easily expressed in terms of the function itself, ie

$$g'(h) = 2\beta g(1-g) \quad \text{for eq 2.49} \tag{2.51}$$

$$\text{and} \quad g'(h) = \beta(1-g^2) \quad \text{for eq 2.50} \tag{2.52}$$

The derivative in Equation 2.51 reaches a maximum value for $h = 0.5$ and since $0 \leq h \leq 1$, approaches a minimum at 0 or 1. Since the amount of change in a given weight is proportional to this derivative, weights will be changed most for those units that are near to their midrange and not decisively on or off. This feature may well add to the general stability of the system learning. One other feature of this activation function should also be explained. It is impossible for the network to reach target outputs of 0 or 1 without infinitely large weights. Therefore, in a practical learning situation the values of 0.1 and 0.9 are typically used as the targets, taking these values to replace $0, 1$.

2.9.3 Learning Rate and Momentum

The learning algorithm requires only that the change in weight is proportional to $\partial E^{\mu}/\partial w$. True gradient descent relies on infinitesimally small steps being taken. In back-propagation the learning rate ε is the constant of proportionality: larger values of ε produce larger weight changes. In a practical situation, as large a learning rate as is possible before the system starts oscillating is desired. Suppression of such oscillation can be assisted by the introduction of a **momentum term** by the following modification to the generalised delta rule:

$$\Delta w_{pq}(n+1) = -\varepsilon \frac{\partial E}{\partial w_{pq}} + \alpha w_{pq}(n) \tag{2.53}$$

where the parameter n indexes the presentation number.

The effect is to give each connection some inertia or momentum so that it tends to change in the direction of the average force it feels instead of oscillating with every little kick. If the network is following a plateau region of the cost surface, then $\partial E/\partial w_{pq}$ will be approximately the same at each step n, so that 2.53 converges to

$$\Delta w_{pq} \approx -\frac{\varepsilon}{1-\alpha} \frac{\partial E}{\partial w_{pq}} \tag{2.54}$$

with a larger, 'effective' learning rate of $\varepsilon/1-\alpha$. In an oscillatory situation, Δw_{pq} responds only with the coefficient ε to instantaneous fluctuations of $\partial E/\partial w_{pq}$. The momentum parameter α must lie between 0 and 1 and the overall effect is to accelerate the long term trend by the factor $\frac{1}{1-\alpha}$ without affecting the oscillations. Typically large values of α (~ 0.9) are used. Experiments by Rumelhart et al. (1986a) showed that the same solutions to problems were obtained when they set $\alpha = 0$ (ie no momentum) and reduced ε, but that the overall learning of the network was much faster with larger values of ε and α.

Since it is not always easy to choose appropriate values of ε and α a priori for a particular problem, and the best values may alter during learning, many authors (eg Cater, 1987, Vogl et al., 1988 and Jacobs, 1988) have suggested the method of **adaptive parameters**. A typical approach is to check the effect that a particular weight update has had on the cost function. If it has decreased it, then it may be possible to increase ε while if the process has overshot, then ε should be decreased. The overall scheme may take the form

$$\Delta\varepsilon = \begin{cases} +a & \text{if } \Delta E < 0 \text{ consistently;} \\ -b\varepsilon & \text{if } \Delta E > 0; \\ 0 & \text{otherwise} \end{cases} \tag{2.55}$$

where ΔE is the change in the cost function and a and b are appropriate constants. The meaning of 'consistently' can be based on the last K steps and when a bad step decreases ε it is often worthwhile undoing the step and setting $\alpha = 0$ until a good step is taken.

This adaptive scheme can be extended to make it even more effective, for example, Jacobs (1988) suggested a learning rate ε_{pq} for each connection pq.

2.10 Variations on back-propagation

With the popularity of the back-propagation of errors algorithm in recent years, plenty of consideration has also been given to possible modifications or extensions to the method. The specific goals of these modifications have been to increase the speed of convergence, the avoidance of local minima and improved generalisation ability. There are many ways of varying the standard algorithm. These include changes in the architecture, the size and makeup of the training set and the update rule. Within this chapter, a mention will be made only of the theory behind adjustments to the update rule. This is done on the grounds of completeness as none of them are implemented in practice in this thesis. Discussions of other possible variations are introduced *in situ* through the optimisation section on the analysis of actual data (Section 4.2).

2.10.1 Cost Functions

The use of the quadratic cost function (Equation 2.32) is certainly not the only possible choice. The only restriction on the cost function $F(\xi_i^\mu, O_i^\mu)$ is that it is differentiable and minimised when the arguments are equal. Carrying out this replacement and calculating the new update rule shows that only the form of Equation 2.32 changes, the other back-propagation equations remain the same.

Solla et al. (1988) demonstrated the effectiveness of the entropic measure

$$E = \sum_{i\mu} \left[\frac{1}{2}(1+\xi_i^\mu) log \frac{1+\xi_i^\mu}{1+O_i^\mu} + \frac{1}{2}(1-\xi_i^\mu) log \frac{1-\xi_i^\mu}{1-O_i^\mu} \right] \tag{2.56}$$

This gives accelerated progress in areas where the cost surface is relatively flat while giving no acceleration, which could lead to overshoot and oscillation, when the cost surface is more sharply curved. Hertz et al. (1991b) gives further technical details on other suggested cost functions with associated references.

2.10.2 Minimisation Procedure

Variations in the minimisation procedure have also been generated as, although gradient descent is a very simple optimisation technique and appropriate to network implementation, it is not particularly efficient.

The most practical alternatives use only first derivative information as computation of higher derivatives does not fit into the framework of back-propagation. Most of these variants combine first derivative information with line searches along selected directions. Examples include steepest descent, conjugate gradient and the quasi-Newton or variable metric method. Again all of these are described in more detail in Hertz et al. (1991b).

Genetic algorithms (GA) offer a completely different approach in that they perform a global search of the weight space without use of gradient information. They are thus less easily fooled by local mimima, nor does the cost function need to be differentiable. Conversely not using gradient information introduces a high computational penalty and so compromises between methods have been suggested (Montana and Davis (1989) and also Goldberg (1989) for GA's in general).

2.11 Necessary Size of Networks

After the questions and issues which were raised in Section 2.5, it remains mainly to place any theoretical framework on the necessary number of layers and units in a neural network which are required to represent a set of functions $F_i(x_k)$. Cybenko (1988) gives the answer: *at most two hidden layers* are required, with arbitrary accuracy being obtained given enough units per layer. Furthermore, provided the function is continuous, then only one hidden layer is required to provide the same ability (Cybenko, 1989). In general the number of required hidden units is not known as it is problem dependent and this places a limit on the exact usefulness of these results. It is also possible that a network with more than two hidden layers may require fewer connections in total. Lapedes and Farber (1987) give a non-rigorous existence type proof for these results. It is based on the the fact that

(i) any 'reasonable' function $F_i(x_k)$ can be represented by a linear combination of localised *bumps* that are each only non-zero in a small region of the domain x_k, and

(ii) such bumps can be constructed with two hidden layers.

The output layer then sums the bumps constructed by the hidden layers to produce the desired function in a manner which is analagous to Fourier analysis. These constructions only show that functions are representable by networks with two hidden layers, it does not guarantee that all are learnable using a learning algorithm.

2.12 Applications of Neural Networks

To conclude this introductory chapter on neural network computation, some general applications will be discussed. These include applications which have been used as benchmark tests for finding the limitations of these methods. These techniques have also found their way into the astronomy and physics communities in several areas just in recent years and so some of the most promising of these uses are mentioned. Omitted for now is a description of the practical implementation of these methods and the particular application to time series analysis. This is held over until the next chapter and placed in context with the specific work of this thesis. The examples discussed

here do not aim to provide an exhaustive list of neural network applications, merely a sample of some interesting or relevant cases.

2.12.1 General Applications

The NETtalk project was aimed at training a network to pronounce English text (Sejnowski and Rosenberg, 1987). The input consisted of consecutive characters from written text with the desired output a phoneme code which could be directed to a speech generator. The network was trained on 1024 words and after 10 training epochs began obtaining intelligible speech. After 50 epochs, a 95% accuracy was obtained. The generalisation ability of the network was then tested on previously unseen words producing a 78% accuracy on their pronunciation. Comparison of NETtalk with DEC-talk, a commercial hand-coded linguistic rule based package, showed DEC-talk to be superior but this has to be tempered by the fact that NETtalk learnt purely by example, while DEC-talk was the result of many years of effort by linguists. This is a commonly met property with neural networks; they are most useful when the problem is not fully understood as rule-based algorithms can outperform them if enough understanding is available.

Neural nets have also been exploited as control systems to reproduce dynamical behaviour. Two examples of this are the "truck backer-upper" problem where a neural net is trained to back-up a truck and trailer to a fixed position regardless of initial position, and a similar problem of using a network controller to navigate a car. The exact details of the latter can be found in Pomerlau (1989). After the network was trained on the correct navigation response to 1200 simulated road images, it was possible for the network to steer a car at 5 km/hr on a winding road. Most of the speed restriction was based on the computational time required by the computer to do a forward pass through the network. It was still faster than other non-network attempts which were made.

Finally, we mention the encoder problem. This problem is generally auto-associative, in that the aim is to find a set of hidden units which map efficiently input to output patterns where the input and output are identical. The number of hidden units is kept small to encorage efficiency. Typically the network is of the form N-M-N where $M < N$ and there are exactly N members of the training set. Although there is no generalisation ability of networks required in this problem, it is often used as a framework for benchmarking as the network can be scaled to any desired size and the degree of difficulty varied with the choice of M. The encoder problem also has practical applications in image compression in, for example, transmitting the complex detail of high-definition television.

2.12.2 Astronomy and Physics Applications

Before concentrating on astronomy applications, a brief mention is made to the work of Lonnblad et al. (1990). In this work, a neural network was trained as an identifier for gluon and quark jets. Much of the study was confined to Monte Carlo techniques for analysing e^+e^- events to detect

gluon or quark jets. In the astronomy community several areas of research were identified for possible applications in a recent review paper by (Miller, 1993). These included adaptive telescope optics, object classification and matching and detector event filtering. These are not reviewed in any further detail here, instead references are given for classifying galaxies (Storrie-Lombardi et al., 1992), adaptive optics systems (Lloyd-Hart et al., 1992), and scheduling of observing time for the Hubble Space Telescope (Johnson and Adorf, 1992), as examples of the work involved. Several areas in astronomy more relevant to this thesis have also been broached and these are now explained more fully.

The first of these applications concerns solar flare forecasting. The proper method for utilising observational solar data to predict the occurrence of solar flares is unknown. This has led to some suggestions that Artificial Intelligence (AI) techniques, and neural networks in particular, might be able to assist in the modelling of these events by consideration of relationships present in historical data. The most complete review of flare forecasting using AI methods has been carried out by Shaw in recent years (see Shaw, 1993a and Shaw, 1993b) although other examples exist (Aso and Ogawa, 1993). Shaw concludes that no AI methods, including a connectionist network, have shown any ability which is superior to that currently achieved by human forecasters. Even in comparison with the most simple models no methods stand out, particularly in the success rate of predicting a positive event which is in general disappointingly low. Other recent initiatives include the analysis of Hard X-ray Burst Events as stochastic events against a background level using neural network models (Conway, 1993).

The field of **solar-terrestrial physics** is one of the most popular current areas of application of neural network techniques, including as it does potential applications of this thesis through the effect of high solar activity on the Earth's atmosphere, with decay of satellite orbits a particular example. Wintoft and Lundstedt (1993) has covered several neural network paradigms and their application in solar-terrestrial physics, while (Lundstedt, 1992) considers the more specific topic of predicting geomagnetic storms using solar input data relating to X-ray flares, coronal mass ejections and coronal holes.

Neural networks have been developed to model the temporal variations of relativistic electrons at geosynchronous orbit (Koons and Gorney, 1991). This has been based on using daily observations of the planetary magnetic index (see Section 5.4.1 for more details). The authors remark on the use of the neural network not only for prediction but as a tool for simulating conditions which are rarely observed in nature. This can be done by varying one input while holding the others constant to observe the predicted effects of the neural net model.

One final important paper is that of Koons and Gorney (1990), in which a commercial neural network package is tested for its abilities to predict sunspot maximum, based on the early behaviour of the solar cycle. This relies on the neural network being able to establish some relationship

between the gradient of the early slope and the final maximum value attained for previous solar cycles. A correlation of this type has been proposed by Wilson et al. (1986). This approach does not adhere to the strict time series analysis which will be used in the following chapters. The prediction of sunspot maximum is important in itself and so an attempt to reproduce this work will be included in Section 6.4 for the purpose of completeness.

2.13 Conclusion

In this chapter an introduction has been given to the basic ideas behind the recent explosion of interest in neural network computation. Since all the research I have carried out in this thesis concentrates on the Back-Propagation of Errors Learning Algorithm in Multi-Layer Feed-Forward Networks, this particular paradigm has been covered in detail. It is hoped that sufficient references have been made to allow the interested reader the opportunity to follow up on alternative approaches fairly readily. With respect to back-propagation, the working of the algorithm has been explained, including its potential and its limitations. The description to date has, however, been theoretical in nature and has not addressed in detail the practical implementation of neural network learning using back-propagation. This aspect of the work is delayed until the following chapter where a more detailed discussion can be given. Specifically, the application of the back-propagation algorithm to analysing time series data and potential problems involved in training networks for prediction purposes will be covered. The final section of this chapter (2.12) did, however, describe some general applications of neural network methods which have been investigated in recent years.

Chapter 3

Practical Application Of Neural Network Methods

3.1 Introduction

The purpose of this chapter is to provide the link between the theory of neural network computation contained in Chapter 2 and how this theory is used to analyse various solar and geomagnetic activity indices (Chapters 4, 5 and 6). Thus we describe the practical implementation of the back-propagation of errors algorithm and the software developed to carry out this research. The structure of the chapter is as follows. In Section 3.2 previous relevant work to this thesis is discussed, providing details of the results already obtained from neural network analysis of time series solar activity data. Following this in Section 3.3 is an account of the formulation for analysing time series solar activity data using neural networks, including a justification for using neural networks for this purpose.

The main part of this chapter is contained in Section 3.4 where a comprehensive description of all of the important software written to carry out the research contained in this thesis. The method of network training is discussed along with the use of trained networks in producing subsequent predictions. The relation of the software written to the back-propagation of errors algorithm (Section 2.9.1) and the McNish and Lincoln method (Section 1.6.1) is also shown. Thereafter Section 3.5 discusses general problems in neural network learning and how they affect the analysis in this thesis. Methods of overcoming these problems will also be discussed. This provides the basis for much of the later analysis of this thesis as the effectiveness of these methods in practice is described in later sections of the results chapters.

3.2 History of Relevant Previous Research

Since the re-discovery of the back-propagation of errors learning algorithm, and the corresponding explosion of interest in it, is very recent, the amount of previous work which directly relates to this thesis is relatively scarce. Specifically four papers have considered the prediction of sunspot number or solar flux using neural networks, namely Weigend et al. (1992), Calvo et al. (1993), Williams (1991) and Koons and Gorney (1990). These papers are by no means the only application of neural networks in astronomy and solar physics in particular. Others are described in Section 2.12.2. They are, however, the only ones of direct relevance to the work attempted here.

In Weigend et al. (1992), the use of neural networks in nonlinear forecasting is considered and the yearly sunspot number is one of the data sets used to test the neural network model. The importance of this work is in the recognition of the problems of *overfitting* and *overtraining* in neural network learning. These issues are described in more detail in Sections 3.5.1 and 3.5.2 respectively where the approach of Weigend et al is covered. Furthermore, this paper considered the difference in approach of *single-* and *multi-step* prediction (see Sections 3.3 and 6.2.2 for explanation of these terms). However, this paper only discussed the yearly sunspot number, paying little attention to how the accuracy of predicted values compared to other methods of predicting the sunspot number. Hence, in this thesis we investigate the differences in multi-step, single-step and also iterative predictions for the smoothed monthly sunspot number (see Section 6.2) while always bearing in mind the level of prediction accuracy already achieved for this index. Further, the method described in Weigend et al for dealing with the problems of overfitting and overtraining are investigated further during the analysis of the smoothed and unsmoothed monthly sunspot number, the smoothed monthly and yearly solar flux and also the geomagnetic *aa* index.

In Calvo et al. (1993), the yearly sunspot number is again the index under study in a similar analysis to that of Weigend et al. The authors discuss reconstructing the attractor dimension of the sunspot number (see Section 3.3 for more detail) as a means of justifying the size of network used to predict the yearly data. Again, however, this thesis, by analysing considerably more data and assessing the prediction accuracy in practical terms goes further than their work . The work of Williams (1991) is important in the use of a test-set of data for preventing overtraining of the networks during training (Section 3.5.2). He considers the monthly solar flux and compares his results with Sargent's method of prediction (Sargent, 1978). In this thesis, we investigate further the use of the test-set method over a range of network architectures used. In addition we use the McNish and Lincoln method of prediction as a comparison model because of its established use by NASA and ESOC for prediction.

We will not discuss Koons and Gorney (1990) further at this point as it is used strictly to predict the sunspot maximum. The paper is discussed further in Section 2.12.2 and again in Section 6.4 where we attempt to reproduce the method and then extend it by introducing a couple of variations.

Previous work relevant to this thesis is thus not extensive in nature although a couple of important ideas have been proposed which we investigate more comprehensively in this thesis. In this category is the recognition of the problems of overfitting and overtraining, which as we will show, result in the overall generalisation ability of a network, and the different styles of prediction such as iterative, single and multi-step approaches. However, these ideas have only previously been implemented in limited cases using one solar activity index. This thesis provides a more complete assessment of their effectiveness. Moreover, the comparison of the different styles of neural network prediction with the McNish and Lincoln method as a benchmark test was not considered in any of this cited work and so is new to this thesis. No previous consideration was given to the existence of possible delay in the predictions, nor has any prediction of geomagnetic data been carried out. These provide other completely novel aspects of my research.

3.3 Analysis of Time Series Data using Neural Networks

In the previous chapter, the methodology behind artificial neural computation using error back-propagation in multi-layer feed-forward networks was established. The solar activity data exist in the form of a full history of values. It is now necessary to consider the specific application of neural networks to time series data. First we discuss briefly theoretical aspects omitted from the previous chapter. The basic premise of time series prediction follows the central theme of classical physics: to predict the future evolution of a system from past measurements of it through the construction of a mathematical model of the system in the form of equations of motion which can be integrated forward in time. This procedure is straightforward in systems with a small number of degrees of freedom. In this case we describe the state of the system mathematically by a point $\mathbf{x}$ in a multi-dimensional space Γ, with the dynamics characterised as the motion of $\mathbf{x}$ in Γ. It runs into trouble, however, in nonlinear systems where there are many degrees of freedom, like a turbulent fluid, the weather, the economy or solar activity. In these cases it is simply not possible to keep track of motion in such a high-dimensional space.

However, studies of the dynamics of apparently chaotic systems with many degrees of freedom have shown that dissipation can reduce the effective number of degrees to a small number. Thus the motion of a system, instead of occurring in the high-dimensional space Γ, is confined to a low dimensional subspace, Γ_A, called an **attractor**. This attractor is often a strange **fractal** object with non-integer dimensionality (Mandelbrot, 1982).

It still remains to identify the coordinates which characterize the attractor. This need not be as difficult as it might appear as it is not crucial how the new variables are specified as long as there are enough of them, and importantly, a set of previous values of the quantity to be predicted is a satisfactory choice. Hence a sequence of m measurements should contain sufficient information to

predict the motion on the attractor, provided m is large enough compared to the dimensionality d. This has been tested and supported numerically (Packard et al., 1980).

Furthermore a rigorous result by Takens (1981) proved that there exists a smooth function of at most 2d+1 past measurements that correctly predicts the future value of the variable in question. Thus in principle the prediction problem for a complex dynamical system whose motion lies on a low dimensional attractor can be reduced to a much simpler problem. What Takens' theorem does not give us is the explicit form of the function that accomplishes the desired extrapolation. It is at this point that the use of neural networks enters the picture. Now in general, a neural network is trained to approximate a functional relationship of the form

$$\underline{y} = \underline{\underline{A}}\underline{x} \tag{3.1}$$

where $\underline{y} \in R^n$, $\underline{x} \in R^m$ and $\underline{\underline{A}}$ is $n \times m$. The network learns this relationship through presentation of a set of P input-output pairs $(\underline{x}_i, \underline{y}_i)$. In this case the values of m and n correspond to the number of units in the input and output layers of the network respectively.

This formulation is also appropriate learning time series data. The input pattern presented to the network is of the general form:

$$x(t),\ x(t-\delta),\ x(t-2\delta),\ \ldots,\ x(t-(m-1)\delta) \tag{3.2}$$

where $x(t)$ is the last known value in the time series and δ is the interval between sample points. The output pattern to be predicted is thus:

$$x(t+\delta),\ \ldots,\ x(t+T) \tag{3.3}$$

where $T = n\delta$ is the time interval up to which the values of the series are required. In this way, the true extrapolation mapping is approximated by a function like Equation 2.29, with parameterisation by the weights and thresholds of the network.

One notable example of this method is contained in the work of Lapedes and Farber (1987) in which neural networks were tested for prediction and modelling of the Mackey-Glass differential delay equation. In practice it was found that traditional back-propagation was very slow and so convergence was accelerated through the use of conjugate gradient methods (see Section 2.10.2). They still found that the predictive accuracy was superior to all other prediction methods available (then).

The inevitable question follows: 'Is this theory relevant to prediction of time series solar activity data?' Recent work by Mundt et al. (1991) and Jenkins (1992) investigated the 'chaotic' nature of the solar cycle. Mundt et al. (1991), in particular, reconstructed the attractor using the method of time-delay coordinates described by Takens (1981) and Packard et al. (1980). They note that the sunspot data is inherently noisy, partly because of the semiquantitative nature of the observations and also because the sunspots themselves are products of turbulent behaviour on

the Sun. Thus they carried out a low-pass filter on the data with a cutoff frequency to 'clean-up' the time series. This does not eliminate all the noise from the data and indeed, since some of the 'noise' in the spectrum is inherent in chaotic systems, it is necessary to model the system. From the results of their investigation, they indicated that the sunspot cycle was indeed chaotic and of low dimension. They emphasised, however, that more investigation into the method of noise reduction was necessary to confirm their conclusions. Taken together, the work of Mundt et al and Lapedes and Farmer provides the justification for analysing the behaviour of the sunspot cycle using neural networks. The very existence of the cycle, with the observed changes in period, amplitude and phase, provides the simplest reason for assuming that the data is not a purely random and noisy system and that there exists an underlying relationship, albeit a probabilistic or statistical one, between the current sunspot number and previous values. It is this relationship which we attempt to train a neural network to represent. Hence we proceed with a full investigation of neural network computation for predicting the future behaviour of solar activity.

In 3.2 and 3.3 the choice of m and n is not pre-determined or strongly suggested by the presence of outside factors. In theory, the number of input units, m, is linked to the dimension of the system. Calvo et al. (1993) used this idea as a basis for choosing the length of input pattern to their network. Significantly in Weigend et al. (1992), the authors show that increasing the number of input units does not degrade the predictive performance of their neural network. Thus the important aspect is that the network has sufficiently many input units. This theory applies particularly to *single-step prediction*, ie $n = 1$, where the network is only being asked to produce the next value of the series. $n > 1$ corresponds to what we term in this thesis *multiple-output prediction*, as the network's output layer corresponds to producing several consecutive steps ahead. In this case the network may require more input information to produce the same level of learning ability. Hence, in this situation, the dependence on m and n translates specifically into the question "How long a history of the time series is required to allow a prediction so far ahead?". Both single- and multiple-output predictive ability is investigated for the various activity indices studied in Chapters 4 and 5. We define now (for later use in Section 6.2.2) *direct multi-step prediction* where only one output unit is required but this output is trained to produce a value in the time series k timesteps beyond the last input value. $k > 1$ is assumed for multi-step prediction since the case $k = 1$ is identical to single-step prediction.

The training set for the artificial neural network is made up from the available past history of the time series data. The maximum number of patterns in the training set depends on two factors. The most obvious is the total amount of available data. The other is linked to the number of units in the input and output layers, ie m and n, since the more data required for each pattern limits the number of different training patterns which can be employed. In fact it has proved to be one of the principal areas of study to examine the choices of m and n, along with the number of patterns

P in the training set and the internal architecture of the network, that yield the most accurate prediction values.

This section has mainly considered the background to predicting time series data using neural networks. Although this discussion has been separated from the general theory, which was explained in Chapter 2, it is important to realise that the conceptual gap between the two is virtually non-existent. As part of the principle of supervised learning, the back-propagation of errors learning algorithm requires the presentation of input patterns for which the correct output is known. Training is then carried out until the network produces the required output on presentation of a set of inputs. For the time series data the principle is exactly the same with the input data made up as in 3.2 and the output as in 3.3. During the training phase all of this data is known and so the network is trained through supervised learning to produce the future values from the previous history. For prediction purposes, only the input data is known and the network calculated output is taken to be the predicted values. This is the overview of the procedure. The sofware developed and the practical aspects of network training are now discussed more fully in the following sections.

3.4 Software Development

All of the analysis and results in this thesis were obtained from computer programs written and developed by me. Although commercial neural network packages do exist, and have been used in some previous work (for example Koons and Gorney, 1990), it was not thought appropriate to use one for this research. The reasons for this were that since a lot of the emphasis in the work falls on investigating the effect of various parameters in neural network learning, then personal development of a code for this research allowed a much better understanding of the methods to be built up. Furthermore, since different styles of prediction were to be tried, it was simplest to develop a code which could very easily be adapted to carry out all the requirements of my research. In practical terms as well, owing to the very strong computational nature of this work, it was important for the software to run silently once the initial parameters had been decided. This also allowed many different setups to be investigated concurrently.

There are four main aspects of this research for which software has been written. These are neural network training, predicting from trained networks, the McNish and Lincoln method and a program to evaluate the predictive accuracy achieved. We now discuss this software in more detail, demonstrating how it relates back to the back-propagation of errors algorithm. I use flow diagrams where necessary to illustrate the structure more clearly.

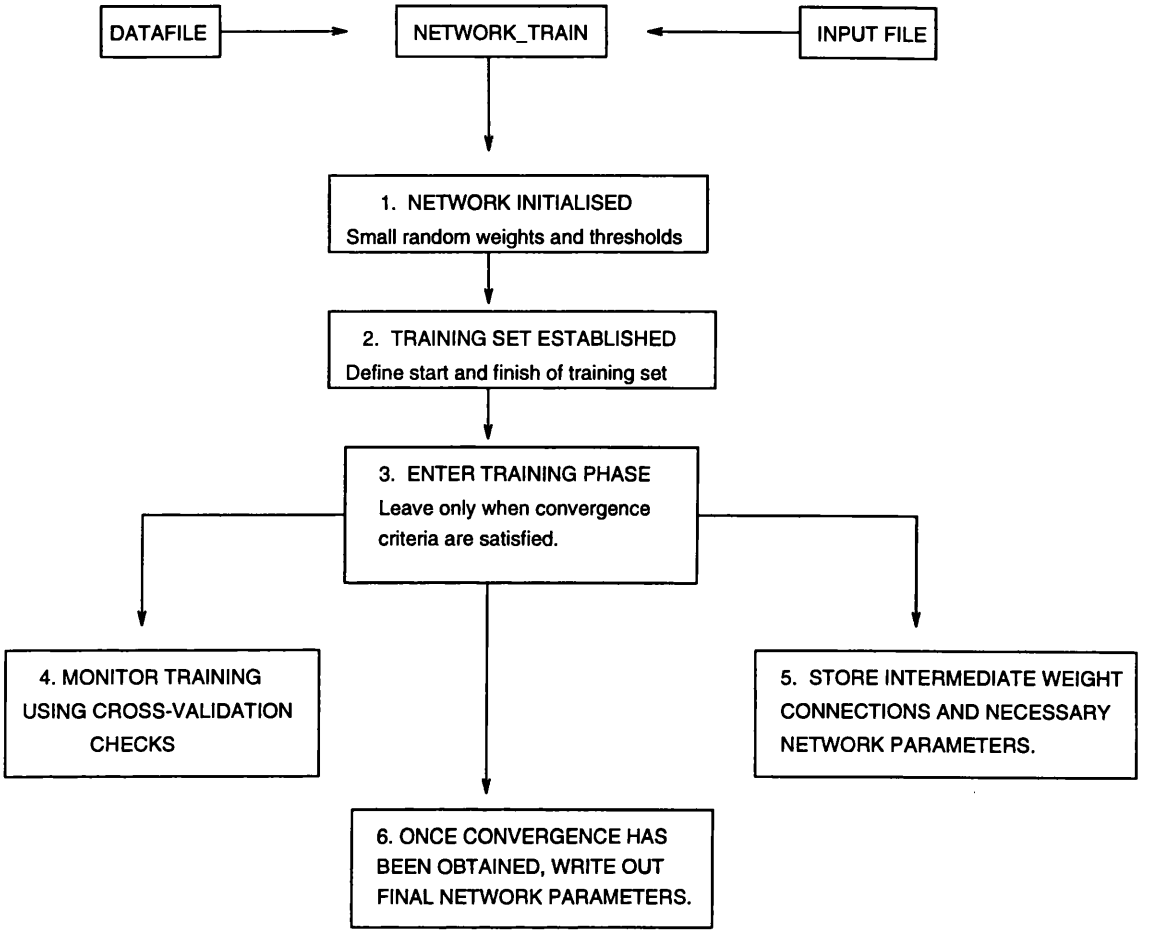


Figure 3.1: The basic structure and layout of NETWORK_TRAIN, a program to initialise and train a multi-layer feed-forward neural network using the back-propagation of errors learning algorithm. The individual aspects are described in the accompanying text.

3.4.1 Program for Network Training

We consider first the program written to carry out the training phase. The basic structure of this program, NETWORK_TRAIN is shown in Figure 3.1. This is for a multi-layer feed-forward architecture learning through back-propagation of errors (as described in Section 2.9). This program takes two arguments, namely a datafile containing the time series data of the solar activity index under study, for example monthly sunspot or solar flux values, and a network input file. The minimum and maximum target values used for the sigmoidal activation function are 0.1 and 0.9 respectively. The reason for this was explained in Section 2.9.2. Thus when the data is read in from the file, it is linearly scaled so that its minimum and maximum values correspond to these limits. Following the prediction phase the outputs will be scaled back relative to the original data. The network input file contains the relevant details about the network architecture and learning

parameters which are to be tested. The various stages of the program are now described in order.

1. The network is initialised using the information in the input file. This specifies the number of layers in the network, the number of units in each individual layer and whether thresholds have to be used on the hidden units. The individual weight connections between every unit in a layer and the next layer (as required by the feed-forward constraint) are then set randomly. By general convention, the random values are small and take positive or negative values. Similarly if thresholds have been chosen, these are also randomly set.

2. The training set for the network is then established. The range of data to be used is contained in the input file and these values are then read in from the datafile. Given the network architecture, the length of the training set determines the number of training patterns. The interdependence of these factors is not entirely obvious, however, depending on precisely how the different input patterns are presented. For this reason, we concentrate on this aspect in Section 3.4.5. For now we assume that the length of training set has determined the number of training patterns. Each pattern consists of a set of input values as in 3.2 with corresponding *observed* output values as in 3.3 (since all of the training data must be known a priori).

3. Following the initialisation of the network, the program then enters the learning phase which constitutes the most substantial part. The set of input values from the first pattern are read into the network input layer, the ξ_k's ($k = 1, \ldots, N_{in}$). These values are passed through the network according to Equations 2.26 to 2.29. Since at this point the weights and thresholds are random, the network outputs $O_i, i = 1, \ldots, N_{out}$ are also essentially random. According to the back-propagation algorithm the error function is constructed as in Equation 2.32. The program then calculates the appropriate modifications to the weights as shown in Equations 2.45, 2.46 and 2.47. At this point the choice is made between incremental updating and the strict gradient descent approach of Equation 2.48. In the strict implementation, often called batch mode updating, the individual weight changes from each individual pattern are summed before the weights are changed once after all the patterns have been presented. Incremental updating makes the modifications to the weights immediately after presentation of the pattern. Experiments by Rumelhart et al. (1986a) showed that incremental updating can be beneficial to network learning and so this approach was used in all the subsequent work in this thesis. Incremental updating also requires no local storage and so in practical terms takes up less memory when the program runs. However, the software was written to carry out both, the choice being made from an extra variable in the input file. Short trials were carried out on the software using small architectures to learn sine curves and no significant difference was found between either method. Thus following the incremental approach, once

the required weight modifications have been calculated, the weights are updated according to these changes. The next set of input values is then read in and the process repeated. In either case of incremental or batch mode updating, we say *one* iteration has taken place once *all* training patterns have been presented. Training continues in this way until certain convergence criteria are satisfied. These are discussed in Point 6 below.

4. During the learning phase, the program enters two major subroutines at regular intervals. The first carries out the cross validation of the network learning. This was done after every hundreth iteration. The aim of this is to assess the generalisation ability of the network and overcome the problems caused by overfitting and overtraining (see Sections 3.5.1 and3.5.2). The different styles of cross-validation are also explained in these sections. The input file determines which form of cross-validation is carried out and sets up the appropriate subsets of data. The cross-validation error is calculated as described in Section 3.5.2 and written to an output file along with the corresponding error on the training set and the number of iterations completed.

5. Corresponding to every cross-validation check is another subroutine which stores all the necessary network parameters as they stand at that point in training. This includes the network architecture, the training parameters such as learning rate and momentum, as well as all the individual weight connections and hidden unit thresholds. Essentially these intermediate files contain all necessary information to be read into a prediction program (see Section 3.4.2 and allow predictions to be made.

6. The final aspect of this program is determining when training should cease. Two convergence criteria were checked after each complete iteration. One to make sure a specified maximum number of iterations had not taken place and another to check the size of the largest error over all training patterns. As we will show with later results, the training error is not the definitive measure of network performance. This is because it takes no account of generalisation error. This aspect was discussed in Point 4 above and Sections 3.5.1 and 3.5.2 and demonstrated in results Sections 4.3 and 5.4.4. Since part of the purpose of this research was to use this software as a test of the cross-validation methods, in practice the programs were run for a set number of iterations with a very strict limit required on the largest error. In this way the program exits the training phase after the predetermined number of iterations. Following exit the final network parameters are stored in a datafile ready for analysis.

It should be noted that while separate programs are discussed for training and prediction purposes, it would be perfectly possible with minor modifications to provide one complete program which did both. The additional subroutines from the prediction program could be added in quite simply and called during the cross-validation checks and/or at the end of training. In addition a variant of

this program was written to allow a partially trained network to resume its training. This simply requires the stored network to be read in instead of the random initialisation stage, and then the learning phase is implemented as before.

3.4.2 Program for Network Prediction

We now discuss the structure of the software developed to carry out the prediction phase. In theory these predictions should be of future values of the time series which are currently unknown. This is after all the particular aim involved in developing the technique. In practice, however, the predictions made were initially over already known parts of the time series so that the effectiveness and accuracy of the technique could be measured. Only after this is done can the technique be genuinely applied to predicting future values of the data under analysis. As was stated in the previous section, the prediction phase could have been written as an extension of the training program. For convenience, however, it was found better to train networks separately, store the intermediate files containing the network parameters and then analyse the results at a later stage. For this reason, the program NETWORK_PREDICT was kept independent. The general structure of this program is shown in Figure 3.2 and follows a similar pattern to that for training. In terms of length of code it is significantly shorter as it does not require the subroutine to carry out the weight changes according to the back-propagation algorithm.

Again this program is run with two arguments, namely the datafile and a network file. The datafile is generally the same as that used for training but the predcitions will be carried out on a different part of it. The style of prediction implemented (see Point 2 below) will affect the amount of data required from this file. The network input file contains all the weight connections and training parameters from a partially or fully trained network. As said before, it contains all the necessary information to recreate the structure and behaviour of the network as it was during the training phase. The individual aspects of the software are now discussed.

1. Firstly the prediction set is established. This could be done either as additional arguments to the program or, since a large proportion of this research was concerned with prediction accuracy over the same range, contained within the program code. The only essential constraint on the prediction set is that it is non-overlapping with the training set.

2. Next the particular style of prediction is chosen. As shown in Figure 3.2 the types of approach listed include single- and multi-step prediction, multiple output networks and iterative approaches. In this context, iterative takes on a different meaning to that of one training iteration. To save on unnecessary repetition later, the differences in these approaches are not explained here but left until Sections 4.1.1, 6.2.2 and 6.2 where the results obtained using these methods are presented.

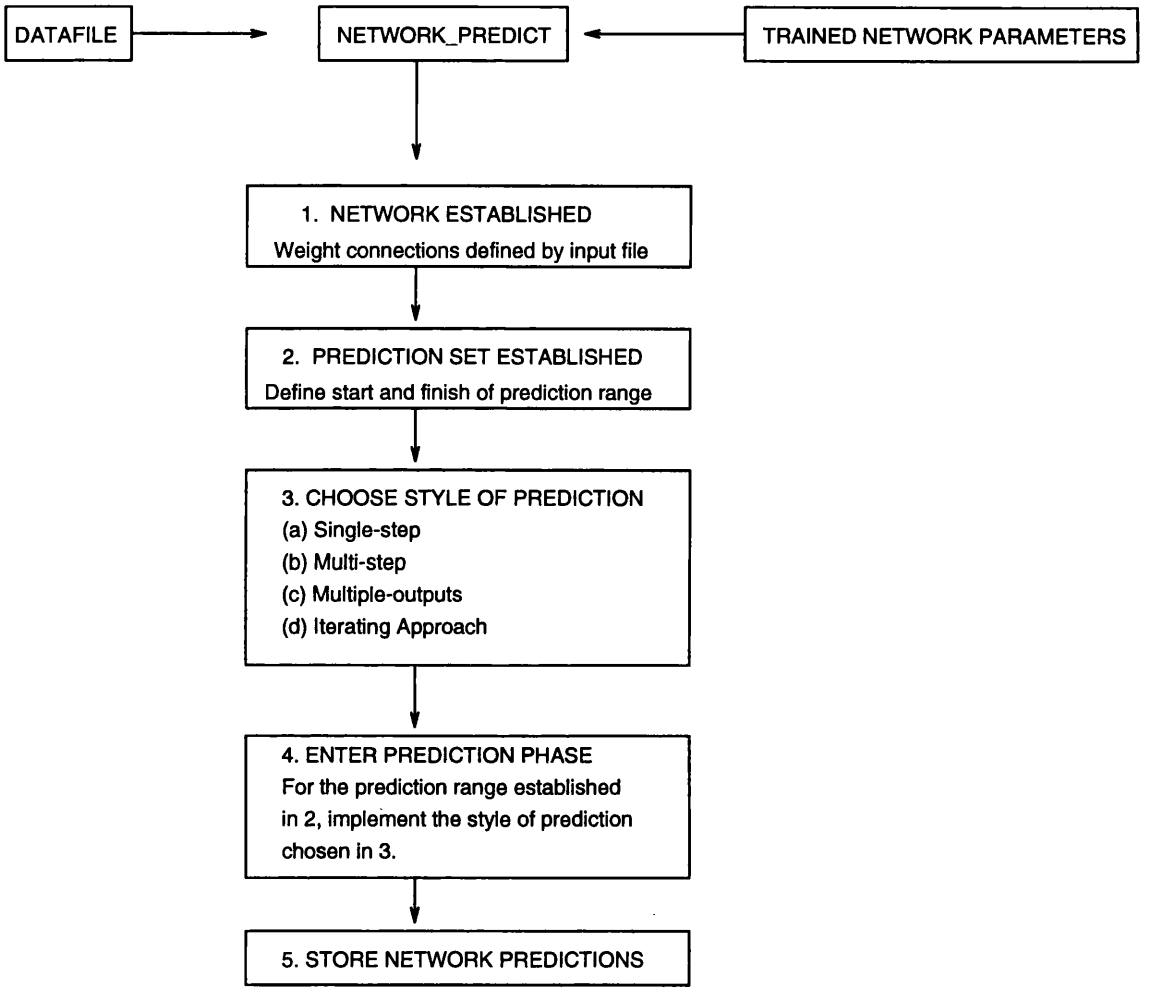


Figure 3.2: The structure of program NETWORK_PREDICT. This program takes a previously trained neural network and uses it to predict other segments of the data. The accompanying text explains the individual parts in more detail.

3. Once the prediction set and style of prediction have been established, the program implements these decisions. The network input file has established a copy of the network structure. The first set of input values from the start of the prediction set is read in to make up the ξ_k's in the input layer. These values are again just propagated forwards through the network by means of Equations 2.26 to 2.29 to generate the output values. Since we are no longer in the training phase these outputs are no longer 'supervised' and instead make up the network's predictions of what the output should be to correspond with the input pattern. The key point comes from assuming that if the network has not been presented with the same input values before, then it can generalise from what it has learnt to produce the correct output. This phase continues until all the input patterns making up the prediction range have been presented to the network.

4. Finally, after each set of predictions have been produced and rescaled back up from between their 0.1 to 0.9 limits, they are stored in a results file. From there, they can be compared with the actual observed values from the corresponding period and the accuracy assessed using various error measures (see Section 3.4.7).

3.4.3 Programs for ML Method and Calculation of Errors

Since the McNish and Lincoln method was chosen as the benchmark comparison test for the accuracy of the neural networks, it was necessary to write a program to simulate this technique as well. The theory of the method was explained in Section 1.6.1. A flow-chart for the program MCNISH.c is shown in Figure 3.3. We do not intend to repeat all the explanation of the workings of this method. Instead the individual features of the program are explained in brief.

1. After the data has been read in to the program, the range of data to be used to calculate the mean cycle is specified.

2. A prediction range is also established, non-overlapping with the data in the mean cycle. This prediction set covered the exact same range as that used for the neural network predictions.

3. Following the procedure summarised in the flow chart and explained in Section 1.6.1, the mean cycle is calculated. Examples of the mean cycle calculated for smoothed and unsmoothed monthly sunspot number were shown in Figure 1.6.

4. For the specified prediction set, the information from the mean cycle is used to calculate the McNish and Lincoln predicted values through the use of Equations 1.7 and 1.8.

5. As before, the predicted values are written to a results file to be compared with the actual data covering the prediction set.

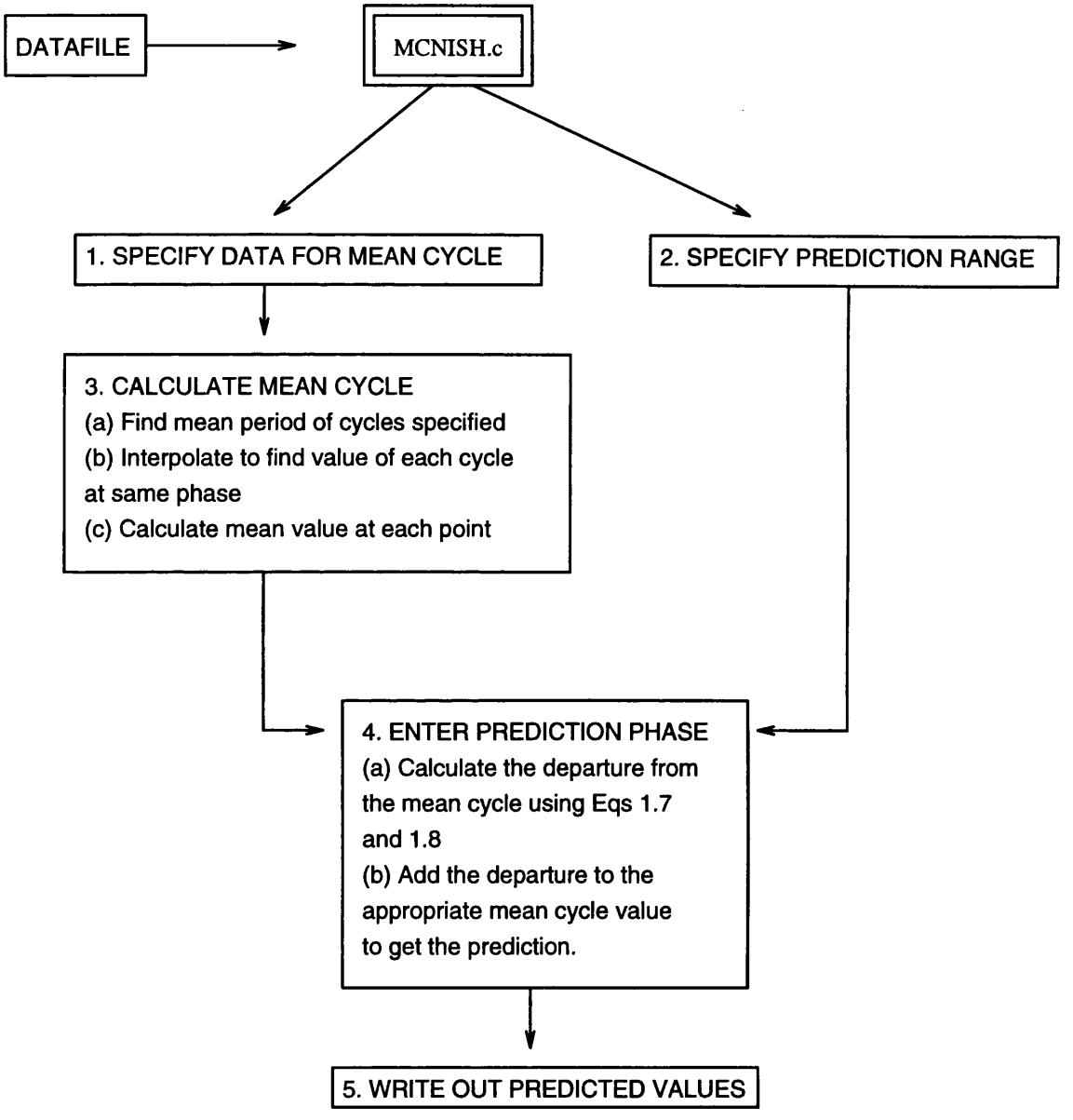


Figure 3.3: A flow chart for the structure of the program MCNISH.c, written to follow the McNish and Lincoln method.

Since an essential part of this research is the comparison of accuracy in prediction of the neural networks and the ML method, a program was developed to calculate various error measures. The criteria for success used are discussed in Section 3.4.7 and so a statistics program was written to calculate the RMS and Chi-square errors. These statistics make up the basis for the majority of the results and conclusions of the later chapters.

3.4.4 Summary of Software Developed

The previous descriptions in this section cover the various programs written in the course of this research. I have attempted to explain the general workings of these programs rather than the specific application. This is because it is important to realise that these programs are not unique to predicting the time series data. The core of the training and prediction programs for neural networks implement the back-propagation of errors algorithm. Thus it is important only to set up the training set appropriately and then any set of input-output pairs can be presented to the network. Details of the actual training sets and the network parameters actually used for analysing the sunspot number are contained first in Section 4.1.1 and then again when the other indices are studied in later chapters. Before that, in the remainder of this section, the precise form of presentation of the training data is discussed again to cover some of the points which were omitted from the discussion of the training software. We will also demonstrate briefly, through experiments on learning sine curves, why the formulation established to learn time series data allows the networks the ability to generalise.

3.4.5 Presentation of Training Data

The explanation of network training above left out some of the precise points involved. This supplementary section aims to redress these omissions. It was stated previously that the range of training data to be used is specified. The main question to be asked is how many training patterns, ie input-output pairs, does this correspond to. For this discussion, we assume that the specified training range has N data points in total and the neural network has m input units and n output units. The number of patterns, P, depends on all of these factors. The first training pattern is made up of the first m datapoints as input values, $x_1, x_2, \ldots, x_m$ say, and the next n points $(x_{m+1}, \ldots, x_{m+n})$ used to 'supervise' the network's calculated outputs. The next pattern is generated by starting at x_2 up until x_{m+n+1}. Thus it is important to note that each point in the training set is *not* unique to only one training pattern. This obviously raises the question of whether this introduces a redundancy in the training data by re-using points repeatedly. We answer this through consideration of the two extreme situations. The maximum and minimum number of patterns are, respectively, when

$$m = n = 1 \tag{3.4}$$

$$m + n = N \tag{3.5}$$

In the first case we get $N-1$ patterns and it is obvious that skipping every alternate point so that no datapoint appeared in more than one pattern would force half of the information to be lost. In the second case there is no choice of patterns but we relax the restriction slightly to consider the case $m+n = N-1$. Here we can have 2 patterns if they are allowed to have points in common. If we assume that $N \gg 1$ then obviously these two patterns would consist mainly of non-unique points and would therefore not be much different from each other. In this case the restriction of one point being unique to a pattern could perhaps be justified. However, the worst that is likely to happen is reinforcement of the strength of the pattern which will not be detrimental to the network learning. This is the assumption behind the formulation of the training patterns.
The details of the actual choice of training set and number of patterns used in the sunspot number analysis (and other indices) is omitted until Section 4.1.1. Returning to the program for network training, in pactice it was found more convenient to specify the number of patterns and the start of the training range as, combined with the network architecture, this defines the range of the training set. As a final point in this area, all the analysis was carried out with the individual patterns presented in chronological order. This was simplest since, having specified the starting point, the next pattern justs moves up one datapoint. No experiments were carried out by randomizing the presentation order of patterns. In strict gradient descent this would of course make no difference since the weight changes from all the patterns are summed before being applied. The experiments of Rumelhart et al. (1986a) suggested that incremental updating was beneficial to learning in allowing the weights to start evolving straight from the presentation of the first individual pattern. There is no reason to assume, however, that the network's predictive ability is changed depending on the order the data is presented although it remains open as a small point to investigate.

3.4.6 Function Learning

Although this chapter is devoted to describing the practical implementation of neural network methods, this has so far only been done in a fairly general sense. At this point we digress to discuss some results obtained when the prototype software was being developed. In any problem the output from a neural network can be expressed as a function of all of its inputs. This function is determined by the update rule (see Equation 2.45 and Equation 2.29) and by the weight connections after training. The aim is to show how a network learns to represent a given problem so that this understanding can be used to provide the network with the best chance to generalise the solution. The formulation already discussed involves learning the future values of a time series from the previous behaviour. This assumes that the value of the data at $x(t)$ is some function f of the previous values. In these simple results we have assumed that no noise is present in the data. By constructing a network of the form 1-N-1 (N hidden neurons), however, it is possible to learn a

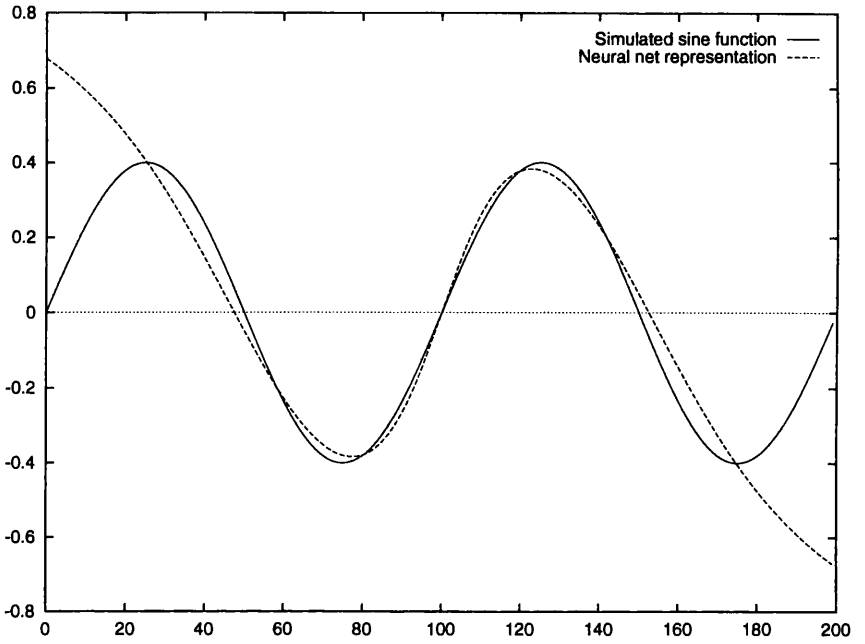
function $f(x)$ as an output given x as an input. In principle this applies also in the case where x is time and $f(x)$ the value of the time series at this time. To test the training and prediction programs written, and also to gain insight into how the network behaves in these different approaches, a simple example was considered. The function to be learned was a sine curve and this was attempted by the following two approaches.

1. A 1-12-1 network trained on $(x, f(x))$ input-output pairs
2. A 5-3-1 network learning a time series generated by $f(x) = \sin(wx)$ in $x \in [0, 1]$

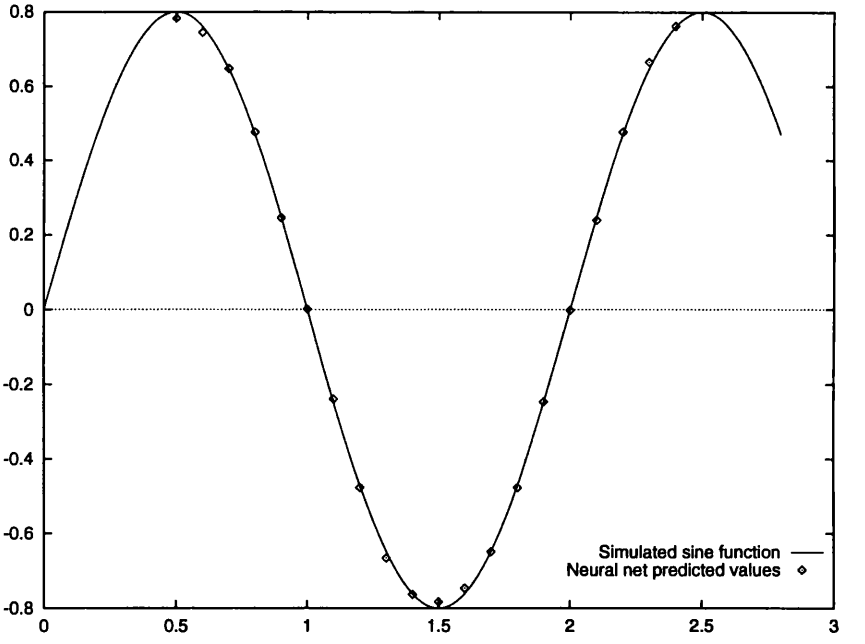
The training phase in both cases is identical. What differs is in the construction of the training patterns. In Case 1 the network has to set up its weights to emulate the function $\sin wt$ in the range [0,1]. The training patterns given to the network consisted of 30 $[x, f(x)]$ pairs. The x value is presented to the input layer and the network is trained to produce the correct $f(x)$ output value. (The network was in fact tested with different numbers of hidden units and training patterns, the latter randomly spaced throughout the training range. More than 12 hidden units allowed equally accurate learning but was slower and unnecessary.) A sufficient number of patterns is important to allow the network to learn the different parts of the sine curve and 30 was found adequate for this. Within the training range, the representation of the function was reasonably successful. This is shown in Figure 3.4(a) where the x-axis has been rescaled with the training range of $x \in [0, 1]$ corresponding to $x \in [50, 150]$. Within these limits the errors on learning are quite low. When it comes to predictions outwith the training interval, however, the network is faced with inputs which are completely different from anything it has previously seen. Thus it is seen in Figure 3.4(a) that the extrapolation outwith the training range of $x \in [50, 150]$ fails completely, suffering from similar problems to polynomial extrapolation.

In Case 2 the 5-3-1 network does not need to approximate $\sin(wt)$, in fact it does not even try to learn it. Instead it learns to predict the next point, given the previous 5. This is free of parameterisation, ie time t is not used explicitly, so the network is not restricted to any domain of t as in case 1. In Figure 3.4(b), the training range was from 0.0 until 2.0 and the extrapolation beyond that is very accurate.

This is an idealised situation of course, particularly as the input pattern starts to repeat. Trials with more complicated combinations of sine curves were tried and the extrapolation aspect, although not as good as this simple case, always showed promise and a far better degree of accuracy than for Case 1. This is because the freedom from parameterisation allows the network a free role when it comes to generalising. Upon presentation of an input pattern of 5 numbers, it can generate an output which is a combination of all other 5 number inputs it has been trained on. This is the justification for implementing the formulation of Equations 3.2 and 3.3. As we will explain in Section 3.5, there are differences in network learning of these simulated cases with small networks



(a)



(b)

Figure 3.4: In the top figure the problem of extrapolating outside the training range ($x \in [50, 150]$) is shown. Freeing the network of explicit parameterisation, allows for much better generalisation capability.

and using larger networks when the relation between the input and output data is not known. For this reason, not too much emphasis was placed on the specific learning capabilities of neural networks in these tests. Their primary use was in testing and debugging the software which had been developed.

3.4.7 Criterion for success

Finally in this section on software development we introduce the criteria of success on which the error statistics program was based. In proceeding with a comparison between the predictive abilities of different neural network architectures, and further, between network predictions and alternative paradigms, it is important to consider in more detail precisely what is meant by good prediction accuracy. We suggest the following criteria:

1. The Root Mean Square Error (RMS) which is defined by

$$rms = \sqrt{\sum_{i=1}^{N}(O_i - P_i)^2 \;/\; N} \tag{3.6}$$

   This is a standard definition and is clearly most applicable to giving an indication of the overall fit of predicted values to a large number of points.

2. The Pearson Chi-Square statistic which is defined by

$$\chi^2 = \sum_{i=1}^{N}(O_i - P_i)^2 \;/\; P_i \tag{3.7}$$

   This statistic is discussed by Holland and Vaughan (1984) as giving a simple and easily applied criterion for judging and comparing goodness of fit of various models to observed data.

3. As a more specific criterion, accuracy in prediction of the maximum value of a cycle and the time in the cycle when the maximum occurs is suggested (see Section 6.4).

Criteria (1) and (2) have both been used in the literature as guides to prediction accuracy of different techniques (see, for example, Mugellesi and Kerridge (1991) and Holland and Vaughan (1984) respectively). When comparing 'network v network', the RMS Error has principally been used to rank the best prediction accuracy. The χ^2 has been used as a backup test and in fact as we shall see, although some minor changes in ordering take place depending on statistic, the general conclusion is not crucially dependent on choice of statistic. For 'network v alternative prediction', the same approach has been used although more consideration is made in Section 6.4.4 of the alternatives available in (3).

3.5 Potential Problems in Neural Network Learning

Before considering the implementation outlined above, it is worth investigating potential problems which may crop up in the training and predictive phases of neural computation. In particular, the problems of "overfitting" and "overtraining" are discussed with respect to the training phase while the possibility of a "delay effect" in the subsequent predictions is also broached. The first two are established features which have been addressed to some extent in previous work (see Section 3.2). The delay effect was not generally recognised by neural network researchers and so the discovery of this effect is one of the original contributions of my work. Since this section deals with problems in learning and prediction, it was thought appropriate to introduce its existence at this time. The evidence of this effect is described in Sections 4.4, 5.2.3 and 5.4.6.

3.5.1 Overfitting

The problem of "overfitting" in the training of neural networks has to be closely studied as it can become very serious in the case of noisy, real-world data of limited recorded length. This is the distinction which must be made following the discussion of the results of Cybenko (1988, 1989) in Section 2.11. Since this has shown that networks can essentially fit any problem (see also White, 1990 for example), in learning computer generated noise-free data no thought has to be given to overfitting. Le Cun et al. (1990) and Weigend et al. (1992) both note the possibility that having too many parameters allows the network to represent not only the underlying signal, but the noise which may also be present. This can have devastating consequences for the generalisation ability of the neural network. It has previously been mentioned (Section 3.3) that having too many input units does not affect the network's performance. This is not contradictory as it is the number of hidden units specifically which provide the network with most of its free parameters. Unnecessary input units can be ignored but too many hidden units connected to necessary inputs can lead to overfitting. Conversely of course, too few connections in the network will not give it the required flexibility to emulate the dynamics of the system ("underfitting").

The work of Le Cun et al. (1990), Weigend et al. (1992) and Hassibi and Stork (1992) represent the recent initiative to provide methods to *prune* a network to the minimal size required to learn a problem. These methods do not pre-determine the size of the network but instead make the network minimise its size as well as the cost function. Le Cun et al. (1990) introduced a technique which they called Optimal Brain Damage (OBD). The basic idea is that it is possible to take a perfectly reasonable network, delete perhaps half the weights and wind up with a network which works just as well. The best generalisation can be obtained by trading off the training error and the network complexity. OBD follows the premise of deleting parameters with small 'saliency', that is those whose deletion will have the least effect on the training error. One of the principal

points of the paper is to indicate that "saliency is not necessarily magnitude". So simply deleting the smallest weights is not appropriate, it is important to identify a true saliency measure. OBD uses second derivative information from the objective function to compute this measure and is validated using a handwritten digit recognition network. Hassibi and Stork (1992) offered an alternative scheme which they called Optimal Brain Surgery (OBS). The basic goal is the same as for OBD but OBS was claimed as a more efficient and effective technique.

Another alternative scheme is that of *weight elimination* (Weigend et al., 1992). In this scheme, the cost function is composed of two terms; the ordinary training error plus some additional term which measures the complexity of the network. It is this function which is then minimised during network training. As part of the demonstration of this method, the authors use the yearly sunspot number as well as currency exchange rates as a test example for theoretical considerations.

At this point in the investigation into neural network models as a predictor for solar activity behaviour, none of the above methods have been explicitly implemented although the reason for their existence has been addressed. Since the main aim of this work has been to find the best network for prediction, the problem has been considered almost by 'brute force', in that a range of network sizes has been tried for each problem. This is discussed in more detail in the relevant section (4.1.1).

3.5.2 Overtraining

A network can become overtrained if it is left learning the training set for too long. In this case the network will have learnt to represent the dynamics of the problem but with subsequent training will start adjusting the weights to fit any noise which exists in the data. Hence although the training error will continue to reduce, the ability of the network to generalise to other data will deteriorate quite fast. This problem is analogous to that of choosing the optimal smoothing function in an inverse problem. The solution to overtraining is to find a way of determining the best stopping point. Two methods in particular have been suggested to try and overcome this problem. Weigend et al. (1992) and Calvo et al. (1993) approach the problem using *internal validation*, while Williams (1991) used a related procedure of a 'test-set' of data. The introduction of the 'test- or check-set' idea for monitoring the network for signs of overtraining and the method of internal validation are motivated by the same reasoning although the execution is slightly different. The underlying principle is to monitor the performance of the network every few iterations against a section of data upon which it is not being trained. In this way the network's ability to generalise can be assessed during training and the problem of overtraining can be tackled. Williams (1991) follows the method of a 'test-set' of data in which a parallel set of input-output patterns, non-overlapping with the training set, is passed through the network during training. For a typical example the error on the training set will be less than on the testing patterns but at some point the difference

in the two will minimise before subsequent training lowers the training error while the testing set level off.

The method of internal validation involves the calculation of the *average relative variance* (arv) on a small section of data from the time series which is not being used for training. This quantity is defined by

$$\text{arv} = \frac{\sum_{i=1}^{M}(O_i - P_i)^2}{\sigma^2 M} \tag{3.8}$$

where M is the number of patterns in the cross-validation set and σ^2 is the variance of the validation set. Again monitoring the behaviour of the arv against iteration number allows network training to be assessed.

Both of these methods do also bear some relation to the overfitting problem. This is because although the number of available parameters in a network is fixed at the beginning, the number of effective parameters increases during training (Weigend et al. (1992) and references therein). The basis for this is that at the beginning of training all the parameters are initialised randomly and small and so weakly-trained networks produce outputs which are very smooth functions of the inputs. As training proceeds, in order to reduce the training error, the weights start to evolve towards specific values leading to more rapid changes in the output for small variations in the input. Continual training thus leads directly to the problem of overfitting as described in the previous section. Thus carrying out some form of internal validation on the network to stop overtraining is also a relevant procedure in dealing with the problem of overfitting.

When applying these methods during the training of neural networks with actual data, the test-set method will be most applicable in the case of sunspot numbers or an extrapolated solar flux dataset. This is due to the fact that this approach relies on having two equally sized datasets for training and testing purposes. Obtaining these while still retaining a large number of training patterns is easier to balance for the longer history of sunspot numbers. It is hoped that the method of internal validation will prove more useful when we are working on geomagnetic indices.

3.5.3 Delay Effect

The potential problems mentioned above relate more specifically to the training stages of neural network analysis. This is a suitable point at which to consider any effects which might show up consistently in the predictions. We consider the existence of the Delay Effect and propose two causes for the problem as well as a means of testing for it. This effect has not previously been recognised in any of the previous work described in Section 3.2 and was discovered through the research in this thesis. We discuss it here prior to the results which showed up its existence simply because this section deals with problems in learning and predicting and provides space to discuss possible causes and consequences without interupting the flow of later results sections.

The delay effect is as simple as it sounds, the predictions of the network being apparently of the

same magnitude as the target values, only time shifted so that they appear to be late. The length of the time shift would be called the delay time. Two potential causes of a delay are *echoing of the inputs* and *maxima* or *turning points* in general. A well documented result of neural network training is that it is possible for the network to learn something other than the exact input-output patterns it is being presented with. In the simple case of predicting one point ahead in a time series, if the difference between two consecutive data points is small, then the network can achieve small errors simply by echoing the input value as the output. Turning points are also possible areas for a delay effect to appear in the predictions. This is based on the intuitive idea that it is harder to track the target values when they are changing in direction rather than monotonically increasing or decreasing. Sunspot maximum is a particular example of this and an important one, as for many purposes the maxima have the most influence on the space environment (see Section 1.5). The first of these effects is most likely to upset the prediction accuracy when the data is relatively smooth and so consecutive points are closer in value. Conversely, a noisy time series which is constantly changing direction will cause more of a delay due to these direction changes. The smoothed and unsmoothed monthly sunspot number provide an example of each of these types of time series and so evidence of any time delay in the predictions will be sought. One method to test for this is to "slide" the predictions backwards along the time axis and calculate an error measure (see Section 3.4.7) between the time-shifted prediction curve and the actual time series value. If the error is minimum at any time-shift other than zero, the data can be said to be suffering from a delay; the delay time being the time-shift corresponding to that minimum. If this delay time is equal to the prediction time then direct echoing could be present in which case the network is not learning to predict the outputs. We use echoing to describe the situation where the network is simply producing the value of the last input as the calculation for the next output. We have found that it is possible for the delay time to be the same as the predict-ahead time without echoing taking place although this is discussed more fully in Section 5.2.3 and 5.4.6. The two effects can be differentiated by 'delaying' the network predictions and comparing the accuracy of those results with those available through reproducing (or echoing) the last input value. In practice, where multiple inputs and outputs are used and the time series is noisy, direct echoing is unlikely but nevertheless may interfere with the 'true' prediction accuracy.

3.6 Conclusions

This purpose of this chapter was to provide the link between the theory of the previous two chapters and the results of the various investigations into neural networks for predicting solar activity indices. In particular, the software which I wrote to carry out these investigations was described and explained in detail. The two principle aspects of neural network research are the

training and prediction phases as both introduce different practical aspects which need to be addressed. In terms of the training phase, the problems such as overfitting and overtraining which occur in real data situations were considered and ways of overcoming them were introduced into the training software. These methods were based in part on previous investigations which have been carried out. The form of the presentation of the training data was also discussed as this has consequences for the networks' learning capabilities. In terms of the prediction phase, the most important problem discussed was the existence of the delay effect. This effect was revealed for the first time through the work of this thesis. It was discussed here as a problem in the prediction phase of using neural network techniques but the evidence for its existence is not encountered until the later results chapters.

Now that the link between the theory and the implementation has been covered, the next three chapters concentrate on various aspects of results. Firstly, prediction of the smoothed monthly susnpot number is covered in depth in Chapter 4. Thereafter Chapter 5 considers the prediction of other activity indices such as the solar flux and geomagnetic data is considered. Lastly Chapter 6 will discuss different styles of predictions and also describes in detail the specific prediction of sunspot maximum.

Chapter 4

Prediction of Smoothed Monthly Sunspot Numbers

4.1 Introduction

The aim of this chapter is to carry out a full demonstration of the theories developed previously and the exact effect certain parameters have on the learning process. The specific solar activity index which is chosen for this demonstration is the smoothed monthly sunspot number. This choice is based on the fact that this index has dominated the literature of prediction methods for analysing solar activity behaviour (see Section 1.6). It also provides a perfect test set of data for neural network prediction, as it is sufficiently large to provide extensive training patterns, without the added complexity (and size) of the past history of daily measurements.

The analysis of the smoothed monthly sunspot number utilises the software which was described in Chapter 3. It is carried out by first making some initial calculations using 'educated' first guesses for some of the relevant parameters. The analysis then proceeds through a series of optimisation steps, all of which are introduced at the appropriate juncture. In addition as mentioned previously it is essential to carry out comparison tests with at least one other prediction method at each stage. The McNish and Lincoln method was introduced in Section 1.6.1 and the results of a comparison with this method are presented in the course of the chapter.

4.1.1 Procedure for Network Training

The interdependence of the length of input pattern with number of available training patterns has been discussed above. In specifying that this section will concentrate on smoothed *monthly* sunspot number prediction, we start by placing an upper limit on the number of months into the future that it would be reasonable to attempt to predict following the input of just one pattern. The

distinction to be made here involves the exclusion from this section of an iterated approach where the network outputs are fed back into the input layer to provide data to continue predicting well ahead of the last *actual* data fed into the network. This iterated style of prediction is held over for discussion until Chapter 6. The upper limit which was chosen was 18 months (which corresponds to $N_{out} \leq 18$, see below for more details) as this provided predicted values over a timescale longer than the resolution afforded by yearly data. It was not considered necessary to continue predictions for 24 months (or more) as this would be tackled by a network trained on yearly data with far fewer output units required. At this point one major assumption is included in this work to provide a basis at which to start the numerical investigations into neural network analysis of this dataset. This assumption is that although the dominant periodicity in the solar cycle is around 11 years, it may not be necessary to have an input pattern (ie number of input units) covering a comparable span of time, for accurate predictions to be made up to the upper limit which has already been fixed. It is undeniable of course, that the networks must be trained on data which covers several examples of the complete solar cycle, that is the total training set length must be much greater than the individual cycle length. However, the training set can be presented in the form of many pieces (subsequently termed patterns) which may individually be less than the cycle length. Thus, the longer term variations are implicitly included in the training set through the availability of a larger set of training patterns. It is also the case that smaller networks naturally require shorter computational time and so a fuller investigation into their properties can be attempted. For this reason attention has been focussed on networks with 6, 12, 18, 24 or 36 input units. The initial investigations are concerned with the best predictions which can be achieved for various time intervals ahead. Specifically we have chosen networks with 1, 3, 6, 12 or 18 units in the output layer. This corresponds to networks with the capability to generate predictions covering the range from 1 to 18 months ahead for this data. Stated explicitly, any network with n output units has a unit in the output layer corresponding to a prediction $1, \ldots, n$ months ahead from the last known value in the input pattern. The problem of 'overfitting' during training has been discussed in detail in Section 3.5.1. This potential problem is approached by training networks with different numbers of hidden units for each combination of input-output pair from the above lists. The number of hidden units has varied between the bare minimum of 1 up to 48 units. One suggested 'rule of thumb' as a starting point for choosing the number of hidden units in a network is 'half the sum of the number of input and output units'. The range of hidden units tried here provides a rough distribution on both sides of this suggestion.

As a reduction on the required computational time at this stage, with the large number of networks to be trained, the training set was chosen to cover only a period of approximately seven complete cycles. This seven cycle span was taken to cover solar cycles 13-19 inclusive as this range gives a good spread of high and low maxima. Also, this shorter coverage of training set for the networks

allows easy investigation into the use of the test-set methods (Section 3.5.2) since other complete 7 cycles of unused data are easily specified. This is consistent with the bounds placed on the training set in the previous paragraph, as the total length is much greater than any one individual solar cycle. At this stage the training parameters were fixed for all architectures to provide a comparison between them. Section 2.9.3 discussed typical values for these parameters. Hence the choices of $\varepsilon = 0.005$ and $\alpha = 0.9$ were made initially. Each network was trained over 800 training patterns for a total of 6000 iterations in the first instance. The training patterns commenced at the same start point at the onset of cycle 13. Due to the different length of input and output patterns required for different architectures, the exact extent of each training set will be slightly different for changes in the values of m and n. These differences will be slight and will not affect the comparisons. Although it might be expected that larger networks will take longer to train and so it would be unfair to compare them on a fixed number of training iterations, in general the greatest reduction in the cost function is gained during the first 100 iterations. This is because the initial weights change from being completely random to providing an approximate fit to the many patterns which are presented. This has been shown to be particularly true of the incremental implementation of Equation 2.48 since the weights start changing after each one of the training patterns is presented. Each iteration thus contains 800 adjustments to the weights, enough to produce a significant decrease in the initially random cost function after only a small number of iterations. Thus after 6000 iterations it was not expected that the different architectures will be showing different cost function errors simply as a result of the larger architecture taking longer to train. Instead a low cost function should indicate that the particular architecture in question is suitable for the prediction problem it is faced with. This will be confirmed later on in Section 4.2.3 It is acknowledged that while other factors, such as the length of training, size of the training set and specific learning and momentum rates, will affect the precise speed and accuracy of learning, this approach has been carried out with the view that such improvements to a network's performance could apply to all architectures. Thus the aim is to identify first the most suitable architecture before 'tweaking' its parameters to optimise its performance.

While no specific training times were noted for any particular network, since in general many networks were being run concurrently, for the largest network used in this survey, a 36-48-18 architecture, the training time on a Sun Sparc 10/30 is of the order of 10 hours of CPU time.

4.1.2 Initial Results

In order to determine the best network architectures for specific prediction time intervals, full calculations were carried out on predicting solar cycles 20, 21 and the start of 22 at sixty evenly spaced points in the history of network training. For a 1 output network an input pattern ending one month prior to the start of cycle 20 is presented to the network and its subsequent output

stored. A one-month timestep through the actual data is then carried out, generating a prediction for each month through the cycles up until the end of 1991. For n> 1 outputs, a similar procedure is implemented although for each presented pattern, there is an output unit corresponding to a prediction between 1 and n months ahead. Storing the predicted values from the same output unit gives a series of predictions which is $x \in 1, \ldots, n$ months ahead of the last known real data. This has provided full results for all predictions ahead from 1 to 18 months for all the combinations of networks which were trained. In addition this allows an investigation into the validity of the test set method of monitoring the training. Based principally on the rms error, all the results were sorted to provide details of the best predictions obtained on this test. This would specify the most suitable network for each month ahead prediction required and also the number of iterations at which the network could generalise best to previously unseen data. In most cases it was easy to establish a few networks which tended to dominate as having the best prediction success. Since it has also been a primary motivation to compare the neural network predictions with another benchmark method, these initial results are compared with those obtained from the McNish/Lincoln technique (Section 1.6.1). As detailed above, at this stage the training data for the networks only drew from solar cycles 13-19. So as a fair comparison it was decided at first only to compare with ML method also restricted to this seven cycle span for its data. The initial results proved promising enough that they were able to go into a straight comparison with ML using a full knowledge of all previous cycles.

Initially a scatter plot of RMS error against χ^2 error for the complete results set was made to look for any trends. Figure 4.1 contains this diagram. It is clear that three main parts of the plot can be identified. These are the main branch to the lower left, another strong branch to the right of this and the wide scatter to the top right. They are identified as follows. The scattered plot corresponds to the network prediction errors after the first sweep through the randomly initialised weights. Thus the network has not had any specific training and so the random distribution is as would be expected. The lower right branch has been identified as consisting mainly of networks with only one hidden unit. In this case the networks have started learning but reach a point where their restricted architecture means they cannot improve further their representation of a more complex problem. The branch to the lower left has been called the main branch as this contains all the other networks and training times which are contained in the other two categories.

The best initial results for each month ahead prediction between one and eighteen months are summarised in Table 4.1. The best networks have been identified by first sorting the one month ahead predictions with respect to RMS error. On some occasions this clearly identifies a specific architecture, other times two (rarely more) architectures will have the same RMS error to the accuracy quoted. Even when one architecture is indicated the errors at consecutive iteration numbers are fairly close. For this reason the χ^2 statistic has then been used to differentiate between

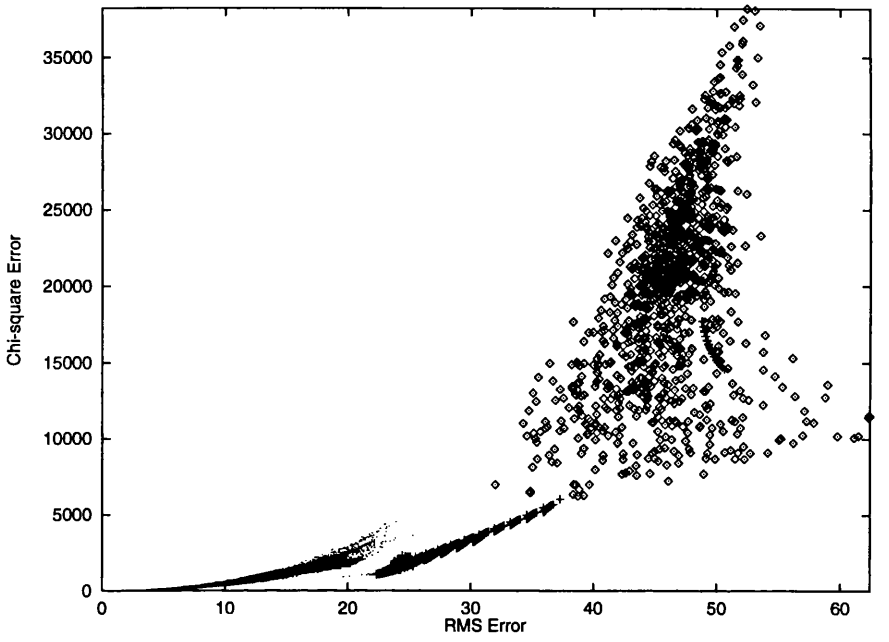


Figure 4.1: A scatter plot of χ^2 v RMS Error for the complete results.

Time ahead	Network	Iteration	RMS Error	χ^2 Error
1	12-18-3	6000	2.4	24
2	12-12-6	4900	3.5	55
3	12-12-6	4800	4.5	104
4	12-12-6	5800	5.7	165
5	12-12-6	5800	6.7	244
6	36-48-12	6000	7.8	282
7	36-48-12	6000	8.5	343
8	36-48-12	5700	9.4	430
9	36-48-12	5500	10.4	536
10	36-48-12	5300	11.3	638
11	24-15-12	3900	12.3	752
12	36-21-18	2500	13.2	845
13	36-21-18	1300	13.8	927
14	36-36-18	500	14.4	869
15	36-36-18	600	14.9	970
16	36-27-18	600	15.4	1024
17	36-48-18	800	15.8	1022
18	36-48-18	800	16.0	1051

Table 4.1: The best initial smoothed monthly sunspot prediction results using neural networks. The specific architecture, training iteration number and error measures are shown for prediction 1 to 18 months ahead.

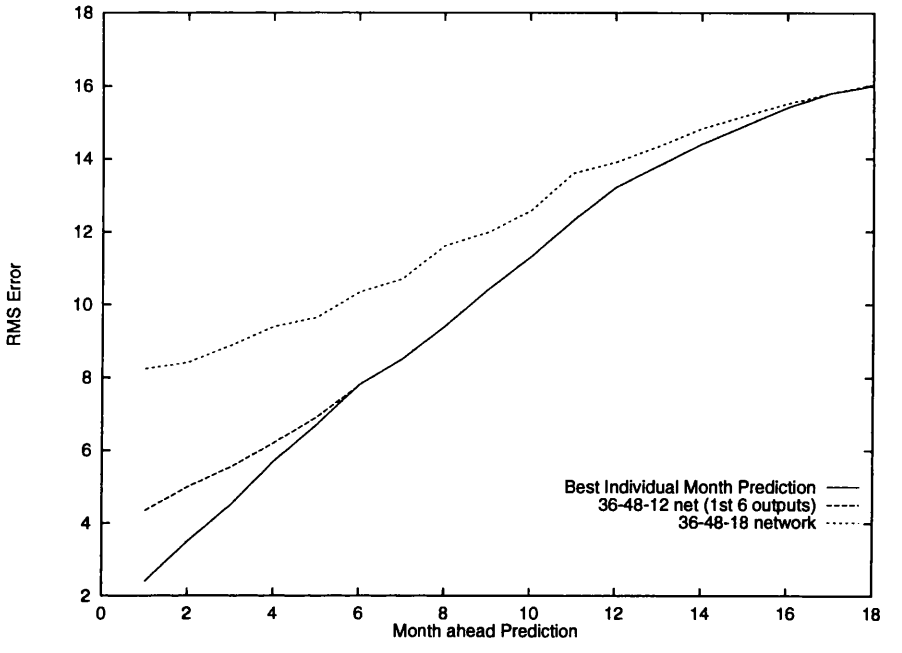


Figure 4.2: A graphical representation of Table 4.1 showing RMS Error against month ahead prediction for the best initial networks. Also shown are the results for each output of the best 6 and 18 month prediction networks up until the intercept with the best overall line.

networks and training iterations. The results of this table are shown graphically in Figure 4.2. Additionally this figure also includes the predictions for the best 6 and 18 ahead networks for their other output units. These shows that the difference in accuracy for the first few months ahead prediction can be significant between the best 1-month ahead network and those for medium to longer term monthly predictions. This makes it impossible to find one network which would be satisfactory as an all round best network between the limits of 1 to 18 month ahead prediction.

4.1.3 Initial Comparison with McNish/Lincoln Predictions

The method initially proposed in McNish and Lincoln (1949) was described in Section 1.6.1 along with the recent variants which have sought to improve upon its performance. The approach of Mugellesi and Kerridge (1991) was included in this description and all references to implementing the ML method refer to a computer simulation of this latter approach which was written for this thesis. Hence this method was used to calculate predictions for 1 to 18 months ahead using first a restricted dataset of solar cycles 13 to 19 inclusive and then using full past history of 1 to 19 inclusive. The predictions covered exactly the same range as for the neural network above. The results obtained are shown in Table 4.2. The comparison is best seen by plotting the results for the best neural network predictions against those for the ML in the previous table. This is done

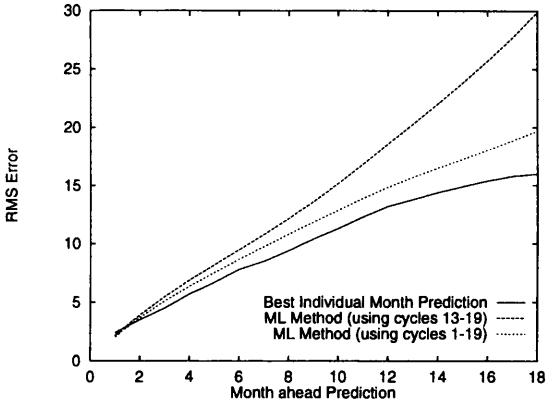
	Cycles 13-19		Cycles 1-19	
Time Ahead	RMS	χ^2	RMS	χ^2
1	2.2	36	2.0	25
2	3.9	105	3.7	69
3	5.5	210	5.1	130
4	6.9	344	6.4	205
5	8.2	509	7.5	292
6	9.5	711	8.7	396
7	10.8	956	9.8	522
8	12.2	1265	10.8	668
9	13.6	1651	11.9	834
10	15.2	2145	12.9	1026
11	16.9	2752	13.9	1243
12	18.6	3446	14.9	1473
13	20.3	4200	15.7	1714
14	22.0	5004	16.5	1956
15	23.8	5866	17.3	2202
16	25.7	6815	18.1	2464
17	27.7	7885	18.9	2731
18	29.9	9062	19.7	3000

Table 4.2: Prediction accuracy for the McNish and Lincoln method over the same prediction set as for the networks. The results are shown when only cycles 13-19 are used and also for the full 1-19 cycles set.

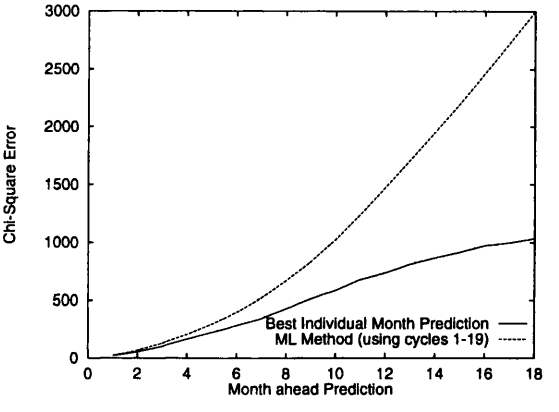
in Figure 4.3(a) using the RMS Error as comparison and Figure 4.3(b) based on the chi-square statistic. From these comparisons we see from Figure 4.3(a) the important point that although on these initial results the best neural network is slightly worse at one month ahead prediction (x = 1), the best network line is from then on lower than both ML predictions, even when the full history of data is used. Figure 4.3(b) shows the same results except using the chi-square statistic where the difference in accuracy grows noticeably as the predictions go further ahead. This confirms the notion that the ML method was only really valid up until about 12 months into the future, although on this test, the neural nets are much more accurate long before this limit is reached. Finally, Figure 4.3(c) confirms the fact that with this data, no single network is able to match the accuracy of the ML method on all the timescales tested here. Instead the 6, 12 and 18 ahead nets (for example) will be worse up until about half their prediction time before they become comparable and then more accurate than the McNish-Lincoln results.

4.2 Optimisation of Results

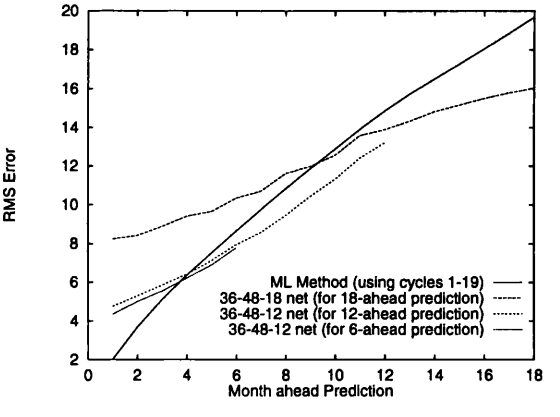
As previously mentioned, the choice of many of the parameters for the initial network analysis was relatively arbitrary, although the results obtained performed well in the above comparisons. The aim of this section is to extend several aspects of the training to optimise the results which can be obtained using neural networks for smoothed monthly sunspot number prediction. By



(a) RMS Error



(b) Chi-Square Error



(c) RMS Error

Figure 4.3: The initial neural network results are plotted against the ML method using firstly the RMS Error and then a χ^2 comparison. Lastly, the best 6, 12 and 18 month prediction nets are plotted against the ML method accuracy.

	Un-optimised			Optimised		
Network	Iteration	RMS	χ^2	Iteration	RMS	χ^2
12-18-3	6000	2.4	24	30400 ($\varepsilon = 0.05$)	1.42	8.5
12-12-6	4400	2.5	23	9500 ($\varepsilon = 0.05$)	2.54	20.8
12-18-6	4300	2.6	26	9400 ($\varepsilon = 0.05$)	2.48	24.6
12-6-3	6000	2.6	26	10000 ($\varepsilon = 0.05$)	1.95	14.4
18-18-1	6000	3.4	47	10000 ($\varepsilon = 0.005$)	2.87	31.7

(a) 1 month ahead networks

	Un-optimised			Optimised		
Network	Iteration	RMS	χ^2	Iteration	RMS	χ^2
36-48-12	6000	7.8	282	6700 ($\varepsilon = 0.005$)	7.77	278.7
36-36-12	6000	7.9	292	6900 ($\varepsilon = 0.005$)	7.86	286.0
12-12-6	6000	7.9	339	5400 ($\varepsilon = 0.05$)	7.83	324.2
24-15-12	6000	8.0	314	10000 ($\varepsilon = 0.05$)	7.65	278.4
36-27-6	6000	8.1	291	16100 ($\varepsilon = 0.005$)	6.98	237.0

(b) 6 month ahead networks

Table 4.3: The best networks with iteration number and error measure for predictions 1 and 6 months ahead. The 18-18-1 network is there for comparison, as being the best 1-output net in the survey. The results under the optimised heading are explained in the corresponding text.

concentrating on 1, 6, 12 and 18 month ahead predictions, the best networks for each were identified. Since the differences in accuracy between the top networks for each regime are small, several networks were identified as worth further investigation. The specific areas to change were (a) using different values of epsilon to attempt to accelerate convergence, (b) train for longer the networks which were still showing improvement at 6000 iterations and (c) extend the amount of training data presented to the networks. The specific 1, 6, 12 and 18 month optimised predictions are now considered separately.

4.2.1 1-month ahead Predictions

Concentrating first on predictions only one month ahead, Table 4.3(a) shows the best prediction accuracy, based on RMS Error, which were found. In this instance it is particularly worth noting that the best 1-output network could not match the accuracy of many of the 3 and 6 output nets. One possible reason for this is that in learning more than one output, the network has to take more account of the direction the output values are going, which provides more ability to generalise than when only one number is required and the network would be more susceptible to noise. All the tabulated network architectures were rerun with several increasing values of learning rate ($\varepsilon[1,2,3] = [0.01, 0.05, 0.1]$) and the networks were trained for a further 4000 iterations in

each case. The 12-12-6 and 12-18-6 had already obtained a minimum and so it was of interest only to see the effect of increasing epsilon for training purposes. The other networks were also continued in their training with the original learning rate until the minimum error was obtained. After a total of 10000 training iterations, the networks were re-assessed. At this point the 12-18-3 network with $\varepsilon = 0.05$ was obviously the best. The RMS error on the prediction set had reduced to 1.75, with the next best being the 12-6-3 architecture with the same ε value with an RMS of 1.95. Since it was still improving over the prediction range, training was resumed to find the best prediction accuracy possible with this network. This network took considerable further training before the network achieved its best accuracy. This was obtained after 30400 iterations with an RMS error of 1.42 and a $\chi^2 = 8.48$ compared with the ML method values of 2.0 and 24.6 for the same error measures. Consideration of the 2 and 3 month ahead predictions available from this well trained network shows a similar increase in prediction accuracy until 30000 iterations. At this point the RMS on 2 month ahead predictions was 2.79 and 3.99 for 3 months ahead. This is now an improvement on the initial results of the 12-12-6 architecture quoted in Table 4.1, and also better than the corresponding ML prediction accuracy of 3.7 and 5.1 (Table 4.2).

4.2.2 6-month ahead Predictions

The spread of best results for 6-month ahead prediction is listed in Table 4.3(b). All of these networks were still improving at the end of the initial training pattern. The same approach as detailed above for carrying out the first stages of optimisation was implemented. In this case changing the learning rate did not turn out to be influential in the learning accuracy of the networks. After 10000 iterations, the 36-27-6 and 24-15-12 nets with the original value of ε had improved more than the other nets. The 36-27-6 network was continued in its training until it achieved a minimum error after 16100 iterations. The RMS error achieved was 6.98 with a χ^2 of 236.8 which when compared with the ML predictions shows a considerable improvement.

4.2.3 12- and 18-month ahead Predictions

Tables 4.4(a) and 4.4(b) show the networks for the 12 and 18 month ahead predictions which were rerun with the different epsilons to see if any improvement could be reached. Since the nets had already reached a minimum, it was not necessary to train them over more iterations. In fact, the results shown in the tables remained the best with none of the other learning rates improving the performance of any architecture. These results provide the confirmation of the earlier statement that an initial fixed 6000 iterations would provide a genuine comparison between cost functions, since in fact all the larger networks had already passed their training minimum within this period. This allays any fears that these networks would require longer training simply due to their size.

Network	Optimised		
	Iteration	RMS	χ^2
36-21-18	2400	13.2	845
36-48-12	4700	13.2	914
24-15-12	3600	13.3	870
36-33-18	2100	13.4	743

(a)

Network	Optimised		
	Iteration	RMS	χ^2
36-48-18	800	16.0	1051
36-24-18	600	16.3	1158
36-27-18	600	16.3	1174
36-36-18	600	16.3	1202

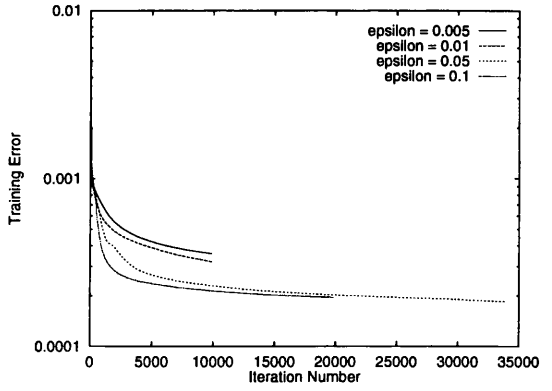
(b)

Table 4.4: The best networks with iteration number and error measure for predictions 12 and 18 months ahead.

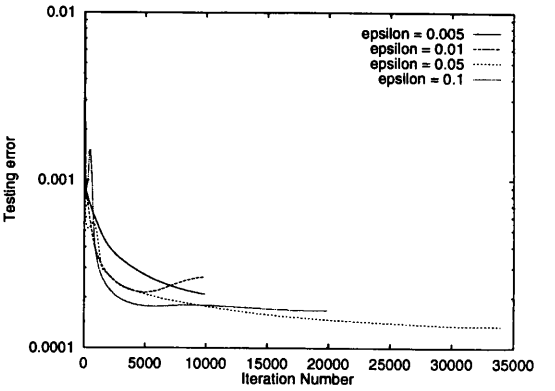
4.2.4 Effect of Varying the Learning Rate

This subsection is aimed at showing in more detail the effect changing the learning rate had on the training, testing and prediction errors. This is done in two cases, first for one of the architectures whose performance was improved and then for a larger network from the 12 or 18 month ahead prediction nets. For the first case, the 12-18-3 network for predicting one month ahead is investigated. Looking at the effect of changing ε on the training error (Figure 4.4(a)), the higher two values show little difference between themselves but are much less at the same stages as the original value. This result, however, can never be conclusive in itself and the effect on the testing error has to be considered in tandem with this result. Figure 4.4(b) shows the testing error behaviour. The original learning rate was showing steady progress but the $\varepsilon = 0.05$ value was lower and also steadily decreasing until it finally levels out around 30000 iterations. Finally, Figure 4.4(c) shows the corresponding effect on the actual prediction accuracy over the specified range. The results there confirm the findings of the first two figures. These results are consistent with the general theory in that the learning rate should not strictly be a parameter of the learning accuracy of a network, provided it is small enough. Thus a smaller value of the learning rate should be able to reach the same prediction accuracy achieved by any larger value, but might take considerably more training time to do so. Ideally, networks could be trained using very small learning rates for a very large number of iterations but this would not reach the balance between desired accuracy and available computational time.

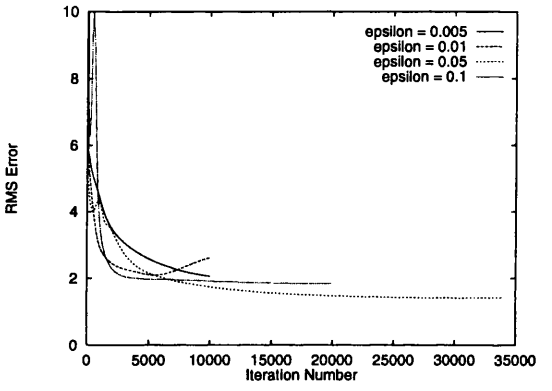
In the case where changing epsilon did not improve the training performance, the 36-48-12 network for predicting twelve months ahead is shown as an example. The training error for the network with the three different learning rates is plotted against iteration number in Figure 4.4(d). In this it is seen that the higher values of ε lead to a greater reduction in the training error early on and remain below the error from the initial learning rate. This would at first indicate that this network might benefit from larger values of ε. Crucially, however, it is the error on the testing set which also has to be considered before any decision could be made. This is plotted in Figure 4.4(e) and



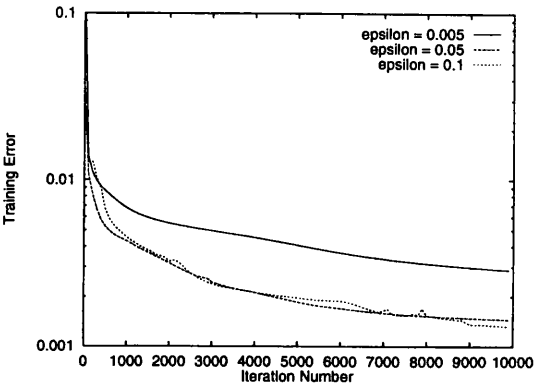
(a) Training



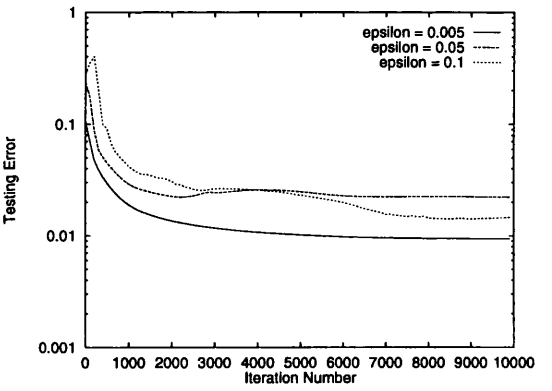
(b) Testing



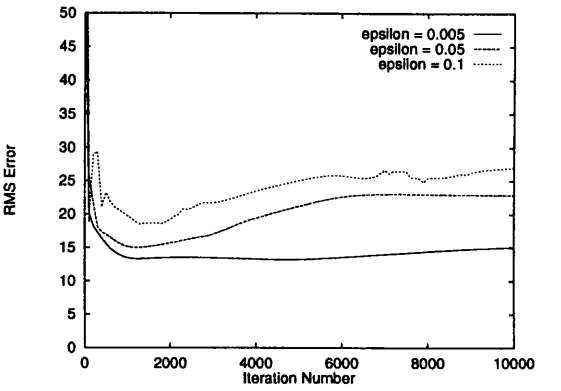
(c) Prediction



(d) Training



(e) Testing



(f) Prediction

Figure 4.4: The effects of changing the learning rate are shown for two cases: the first 3 graphs relate to the 12-18-3 net where the performance was improved through changing ϵ, the last 3 to the 36-48-12 net which showed no improvement in this test.

this time shows the original learning rate considerably better than the other learning schemes. The conclusion which is provided by these graphs is the fact that it is essential to base an evaluation of a network's success not solely on the training error but more importantly on an independent test of the generalisation ability of the network. The result in terms of the prediction accuracy over solar cycles 20, 21 and 22 is shown in Figure 4.4(f) where the accuracy in terms of RMS error for the network with $\varepsilon = 0.005$ is seen to be much better.

These results lead directly to a more detailed study of the effectiveness of the test-set and internal validation approaches for monitoring the training of a network for the best time to cease learning and start using the network for prediction purposes. This is done in Section 4.3 below. Before that the optimisation section is concluded with consideration of extending the training set to take in the full past history of solar activity instead of the 7 cycle subset which has been used up to now.

4.2.5 Training on Full History of Solar Activity

Since in practice this is much more computationally expensive, only an initial investigation has been carried out using a training set made up from smoothed monthly sunspot data from 1795-1962. This left the initial part of the history available for internal validation if required and the most recent solar cycles for actual predictions as before. Two networks were chosen to be the test for this approach. It was appropriate to take the best for one month and six month ahead predictions and extend their training set to look for any improvement in their predictive ability. Following on from the previous section on varying the learning rates, the most suitable value for each architecture was used through training, ie $\varepsilon = 0.05$ for the 12-18-3 net and $\varepsilon = 0.005$ for the 36-27-6 net. After 6000 iterations, the prediction accuracy of these two networks was assessed. This was done simply by comparison with the errors when learning on the shorter dataset. It was found that the 1-month ahead predictions of the 12-18-3 network were noticeably better for the larger training set (an RMS Error of 1.86 as opposed to 2.03), and were still improving. The 2 and 3 month ahead prediction accuracy was however slightly inferior and, more importantly, both had started to degrade in performance by 6000 iterations. None of the predictions for the 36-27-6 net were improved by the use of a larger training set. For the 6 month ahead results, the best was 8.8 at 1200 iterations (as opposed to 9.8 for the original training set at the same stage) but after 6000 this had drifted slightly to 9.0 in contrast to an improvement to 8.1 RMS error. Thus the original network was slower at the initial learning but maintained an improvement in fitting the patterns under study. This might suggest that with the much larger training set, a dependence on the learning rate is possible in the sense that the network would require a smaller ε in order to continue stable learning. The 12-18-3 network was continued in its training and also rerun with a smaller learning rate as the previous section has indicated that the smaller networks are more susceptible to changes in this parameter. The best 1-ahead prediction accuracy obtained was 1.46

after 18100 iterations, a fairly insignificant difference from the original results. The 2 and 3 ahead outputs were also marginally less accurate. This would indicate that for predictions ahead on these timescales, then it is not *necessary* to include the full history of solar data to achieve the same order of prediction accuracy, and use of a limited training set has considerable advantages in terms of computational cost. This might also confirm the thought that the network retains some sense of the global training set as some tests have indicated that a neural network performs best when using a prediction set which immediately follows chronologically the training set (Hertz, 1993a). Thus ultimately it may prove more advantageous to retrain the networks including the most up to date measurements when predictions of the future sunspot numbers are required. These results would also suggest that on the timescales of variation considered here, the Sun itself does not have a long term 'memory' of its past behaviour. This does not rule out the possibility of longer term modulations such as the proposed 88-year Gleissberg cycle which would have to be considered in predictions of solar activity over comparable timescales. Use of the smaller learning rate $\varepsilon = 0.005$ led to results which were poorer than those listed above at the same training stage and were still requiring further training after 30000 iterations.

4.3 Effectiveness of Test-Set and Internal Validation Approaches

The potential importance of these methods for analysing, during training, the corresponding performance in prediction of a neural network has already been implied in Section 4.2.4 on using different learning rates. Having carried out the numerically intensive phase of training many networks of differing architectures the opportunity exists to test the usefulness of these methods in practice when dealing with the real time series data of solar activity. From the table of best initial networks (4.1), we consider first the 12-12-6 network which would be considered the best all round network for predictions 1-6 months ahead. Considering each individual output of this network, they each minimised over the specific prediction interval at a different number of iterations. Since this network was trained on the 7-cycle subset of training data, we study this primarily as a check on the test-set approach.

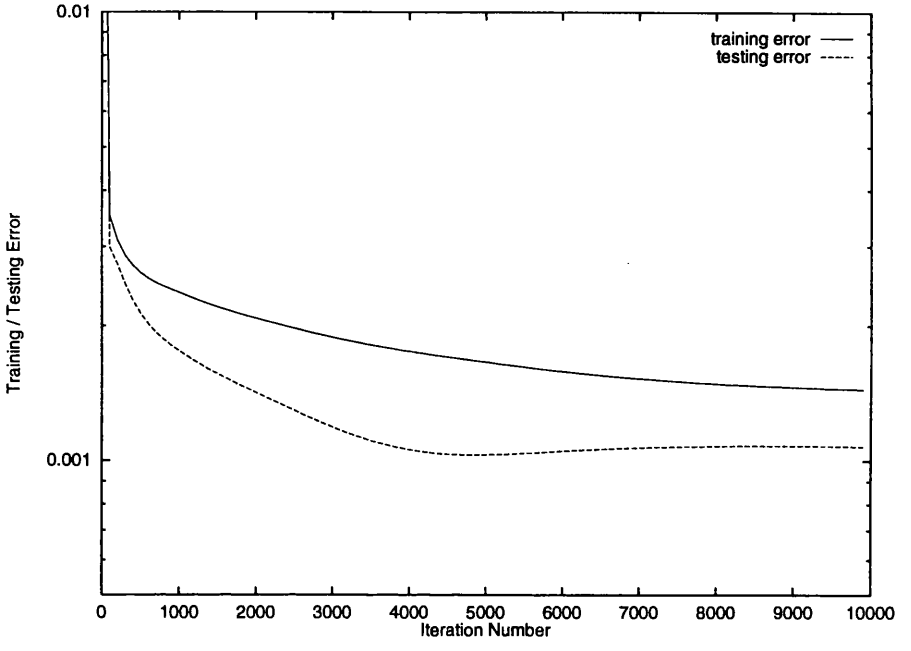
4.3.1 Test-Set Approach

The effectiveness of this approach can be gauged from studying a graph of the training and testing error, as seen in Figure 4.5(a). This graph shows up slightly differently from what was expected in that the error on the testing set is actually less than the training error. This is perhaps surprising since the network is being specifically trained on one set using back-propagation but has no such advantage with the test set. It is merely presented with the test set data, no weight corrections

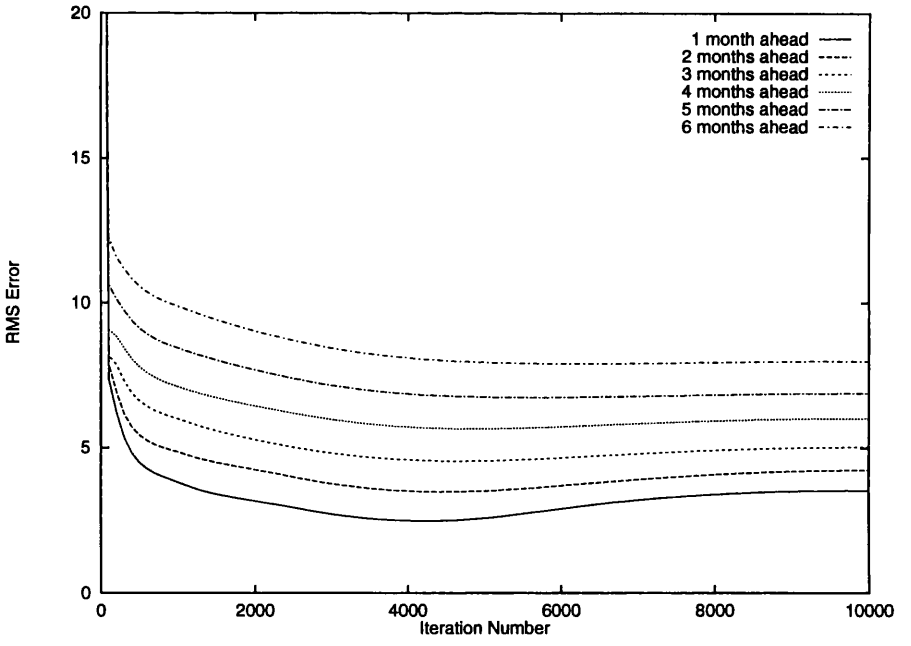
are carried with respect to these presentations. It can be explained reasonably simply by the fact that in general solar activity was slightly lower over the earlier cycles. The form of the error function (Equation 2.19) is designed to learn on magnitude error rather than percentage error. This would allow the magnitude of error on the earlier test set data to be less than on the training data while the percentage error was actually higher. Despite this fact the individual lines on the graph do follow an expected pattern of behaviour. The training error shows a continual decrease in the error and is still slowly improving after 10000 iterations. This shows that the network is still adjusting the weights to fit the given data slightly better at each stage. A glance at the error on the testing set shows up a different impression. After the initial decrease in error, a levelling off occurs around 4000 to 5000 iterations followed by a marginal deterioration thereafter. This indicates a stopping point for training around the minimum as the best point of generalisation, after which the network is starting to *overtrain*. In practice the predictive ability of each of the network's output units is plotted in Figure 4.5(b), with RMS Error against iteration number. The most variation in accuracy is for the 1 and 2 month ahead predictions which both reach a minimum at 4400 iterations. The other curves do not minimise until slightly later but all have reached their best point by 5800 (4 and 5 month ahead curves). Although this is a few hundred iterations after the most obvious suggested stopping point, in practice the difference in the accuracy of predictions for 4 and 5 months ahead is negligible for 5000 and 5800 iterations (6.75 to 6.74 for the 5 month). Thus the effectivness of the test set as a method of monitoring the training to detect the best stopping point is demonstrated. In implementing the approach it may be necessary to train past this point before it becomes obvious but this is at least more desirable than choosing an arbitrary training period or continuing training as the learning is apparently still improving when in fact the predictive ability of the network is rapidly getting worse. Following this more in-depth discussion, the importance of both Figures 4.4(d) and 4.4(e) is further explained.

4.3.2 Internal Validation Approach

Since this method follows a similar thought process to the test set, it is not investigated for the cases above where enough data exists to provide a testing set. Since internal validation requires much less data, we use the results of section 4.2.5 where a much larger training set was used as a check on this approach. Figure 4.6 plots the behaviour of the training error and the internal validation error against iteration number for the 12-18-3 network trained on the full past history of data. The training error shows the expected profile. The internal validation error fluctuates considerably in the first few thousand iterations reaching a first minimum after 1500 with a maximum followed at 5500. Thereafter a more consistent profile is observed with a further minimum around 16500 iterations, closer to the actual minimum of 18100 for this prediction set. Weigend et al. (1992) does observe that one potential flaw in internal validation is the possibility of local minima being



(a)



(b)

Figure 4.5: An investigation into the effectiveness of the test set approach of monitoring training shows the suggested stopping point from the top graph (around 4000-5000 iteration) corresponds to the best generalisation in actual predictions in the bottom figure.

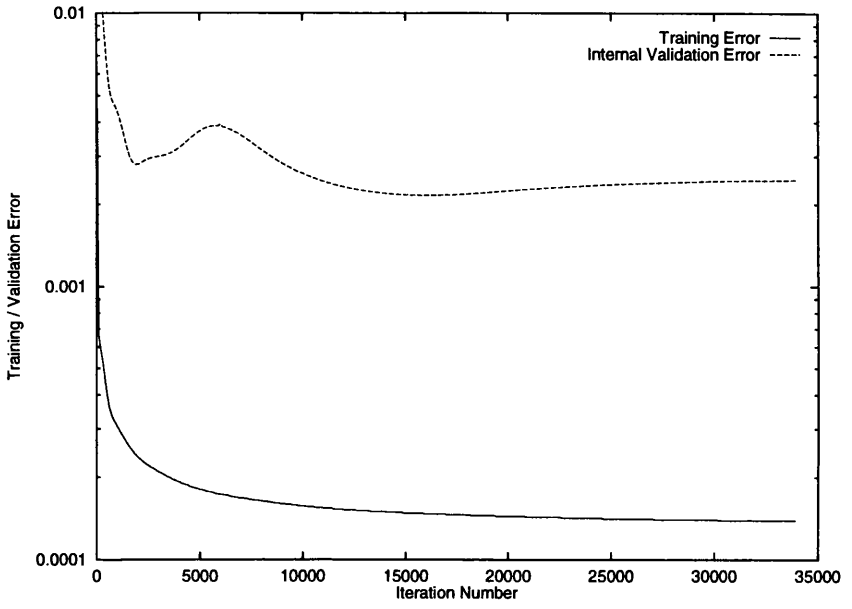


Figure 4.6: A plot of the internal validation error and the training error against iteration number for the 12-18-3 network.

reached before the true value is found. This would seem to be borne out with this result making this approach perhaps less desirable than the test set, but necessary in the case of limited data. An important comment to be made concerns the observed fact that in the results obtained to date, the output units corresponding to predictions longer ahead require fewer training iterations to reach the optimal generalisation point. This has consequences for the implementation of the above approaches which monitor the training over all the outputs. Thus it is probably more constructive, particularly for networks with quite large output layers to be monitored on each output individually. Instead of demonstrating this fact now, it is left to the equivalent analysis section in the next chapter when the geomagnetic indices are under study. In the example used for the check-set because the network studied was the best 'all rounder', the different stopping points all came close to the same time in training and so the global measure was sufficiently accurate for each output unit.

4.4 Evidence of Delay Effect?

The possibility of a delay effect contaminating the prediction accuracy of the neural networks was introduced in Section 3.5.3. A check for the influence of this was suggested at the time, namely to time-shift the network predictions with respect to the actual data, to find a minimum value at a particular shift. This is carried out for the best 1, 3, 6, 12 and 18 month ahead prediction networks.

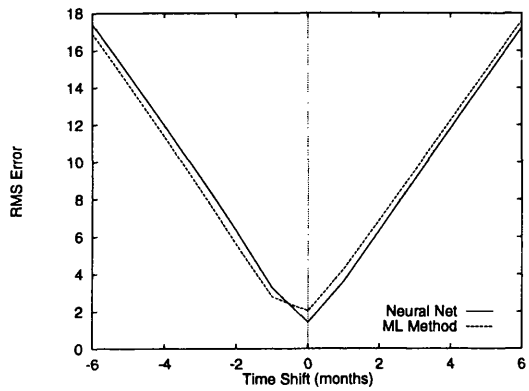
The RMS and χ^2 errors are calculated for a range of time shifts either side of zero. The range is taken to be several months more than the prediction ahead time to give a clear picture of the behaviour. Figures 4.7(a), 4.7(b), 4.7(c), 4.7(d) and 4.7(e) show the variation of RMS error with time shift. The shape of the graphs were the same for the χ^2 error. Looking at these graphs in turn, importantly from Figure 4.7(a), the minimum is at a zero time shift and so no delay (or echoing for this case) is present. For the 3 and 6 month ahead predictions the minimum is slightly offset from zero although by no more than one month in either case. The neural network predictions for 12 and 18 month ahead start to show a more noticeable delay of around 3 and 4 months respectively. This has to be considered in context with the fact that this is not severely restricting prediction accuracy since the time ahead of prediction is much greater than the suggested delay. From all of these figures it is clear that the ML Method does not show any real evidence of a delay for any of the specified predictions here. This is not considered a flaw in the neural network predictions when compared to the ML, the prediction accuracy is already superior, but more particularly, if the delay could be corrected for the neural network results, then a shift of the curves in each of the graphs would result in a bigger difference between the accuracy of the network and the ML predictions. Consideration of possible correcting techniques are not mentioned here but instead are left until Sections 7.3.1 and 7.3.2 for discussion as future avenues of research.

4.5 Conclusions

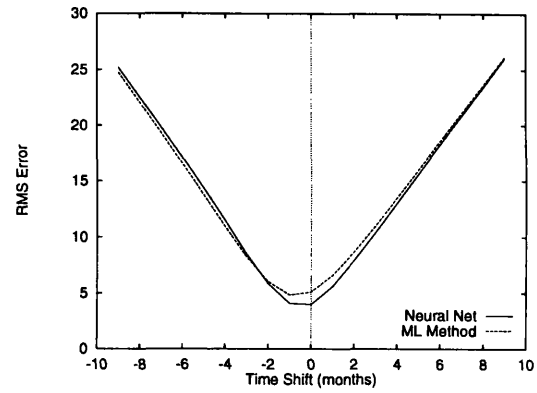
In this concluding section, it remains to summarise the results obtained through the different optimisation sections which have comprised the latter parts of this chapter, in conjunction with some examples showing the predicted values of the sunspot number against the actual behaviour. Following this we discuss the lessons which can be learnt from this initial study as we proceed to the next chapter where we extend the range of indices and timescales for which neural network prediction is sought.

4.5.1 Summary of Prediction Results

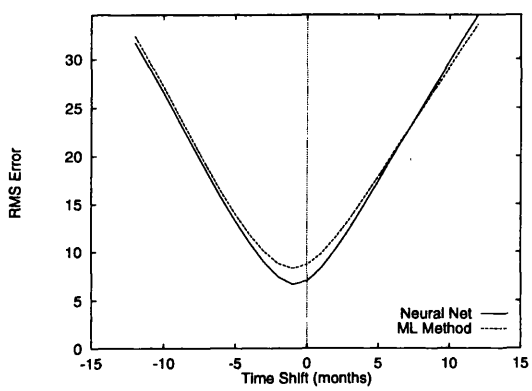
The work of this chapter proceeded through the presentation of the initial results which were obtained from the wide range of network architectures trained with specified parameters. Subsequently, the effect of varying these parameters was considered through the various optimisation sections. Although the results of those sections were described in detail at the time, it remains to draw together the variation of different parameters to provide a final summary of the best prediction accuracy obtained in this work for the smoothed monthly sunspot number. The biggest difference on the initial results came through the variation of the learning rate. It was found that for smaller architectures predicting only a few months ahead, that higher values of the learning



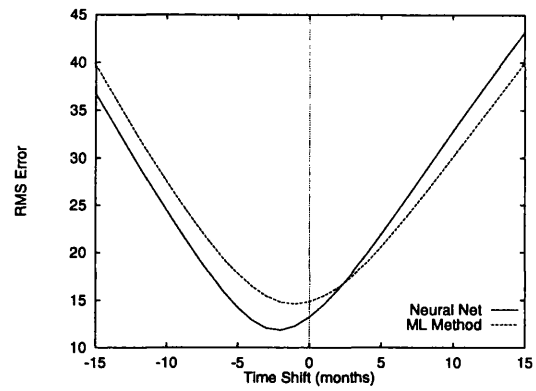
(a) 1 month ahead



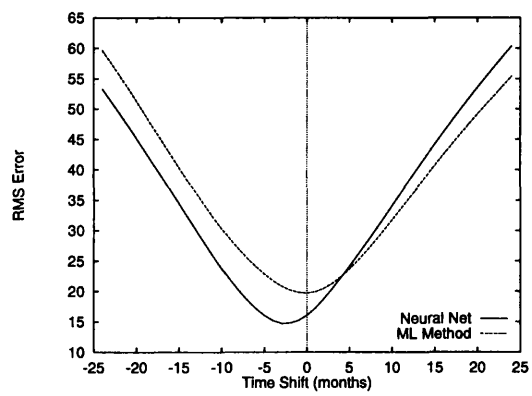
(b) 3 months ahead



(c) 6 months ahead



(d) 12 months ahead



(e) 18 months ahead

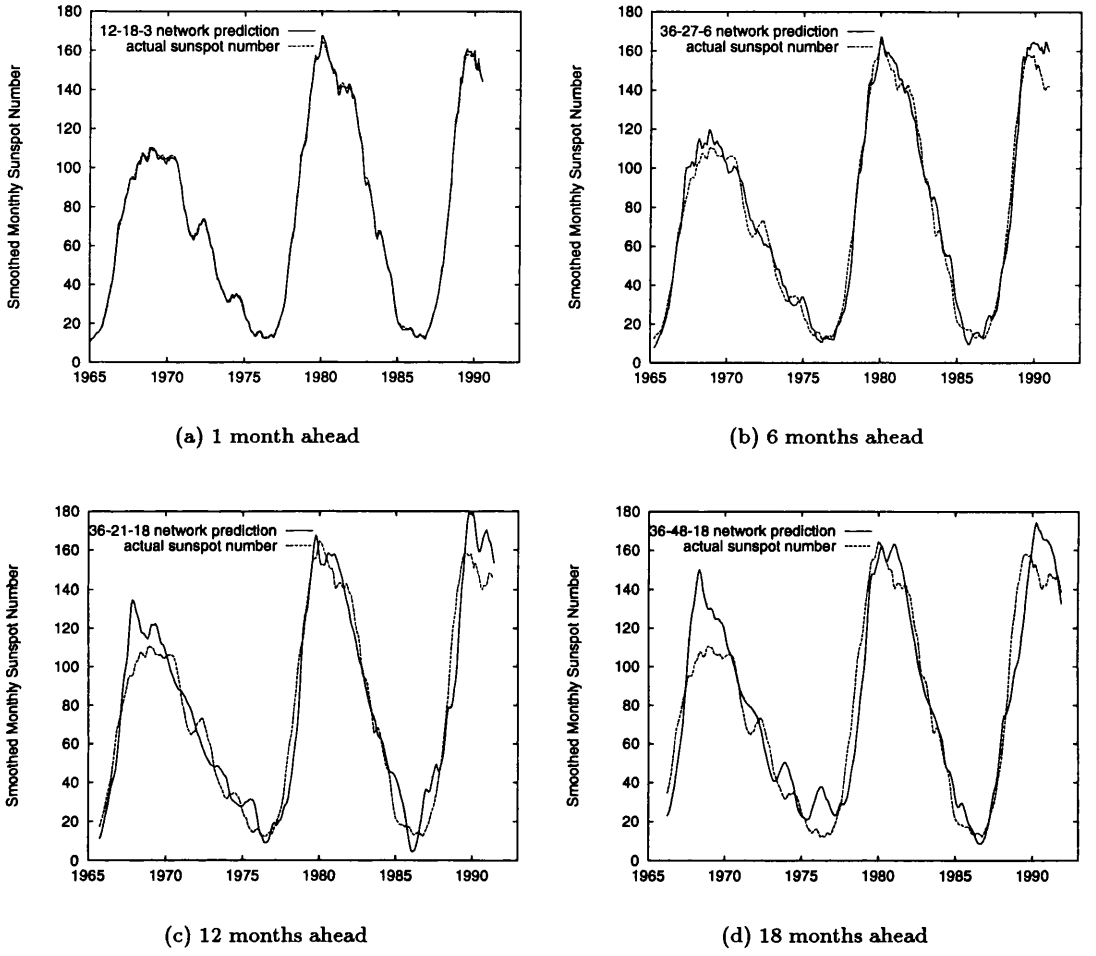
Figure 4.7: Existence of any delay effect in the predcitions is checked for on the specific 1, 3, 6, 12 and 18 month ahead predictions.

Time ahead	Network	Iteration	Learning Rate	RMS	χ^2	ML Rms	ML χ^2
1	12-18-3	30400	0.05	1.42	8.5	2.0	24.6
6	36-27-6	16100	0.005	6.98	237	8.7	396
12	36-21-18	2400	0.005	13.2	845	14.9	1473
18	36-48-18	800	0.005	16.0	1051	19.7	3000

Table 4.5: Summary table showing the optimised predictive accuracy achieved for 1, 6, 12 and 18 month ahead predictions. The network architectures, learning rate and number of training iterations required are listed for each case.

rate than were initially used allowed stable and more rapid learning to take place. The initial values chosen appeared appropriate for the larger architectures as stable learning was achieved but increasing the learning rate showed a general decrease in learning ability. The other main optimisation which was attempted was the extension of the originally limited training set to cover the full past history of available data. The result of this was that it proved unnecessary to attempt the additional computation involved in this optimisation procedure as the prediction accuracy from the shorter set proved just as accurate as the full set. This indicates that all the necessary patterns for learning variations on the timescale of smoothed monthly behaviour are available in the reduced 7-cycle training set originally used.

Using these results, the best prediction accuracy obtained through this study for the particular months ahead which were studied, namely 1, 6, 12 and 18 months, are summarised in Table 4.5. This shows the month ahead prediction, network architecture, the number of training iterations, the learning rate during training and finally the prediction accuracy, quoting also the best results using the McNish and Lincoln method as the comparison measure. Having discussed the prediction capabilities of the various network architectures through the use primarily of the Root Mean Square Error Measure, it remains to demonstrate how these prediction accuracies translate into graphical form showing the predicted monthly sunspot values against the actual behaviour. Figure 4.8 shows the 1, 6, 12 and 18 month ahead predicted values of the sunspot number against the actual sunspot number for the test period. In the case of the one month ahead prediction, the differences between the actual and predicted values are small enough that it would be necessary to look at subsections of the prediction range in order to obtain a clearer picture of the accuracy of the results. In all cases, however, it is in fact more constructive to look at the residuals or errors in the predicted values throughout the solar cycle. The percentage error is also studied as this gives a better understanding of where in a cycle the network does well and where it would need more training. Considering first the case for the one month ahead prediction, Figure 4.9 shows both the sunspot number residuals and the percentage error for the predicted set and also for the last three cycles in the training set. It is important to confirm that the network shows up the same properties when generalising to unseen data as it exhibits on the training set. Figure 4.9(a) confirms that the behaviour is



(a) 1 month ahead

(b) 6 months ahead

(c) 12 months ahead

(d) 18 months ahead

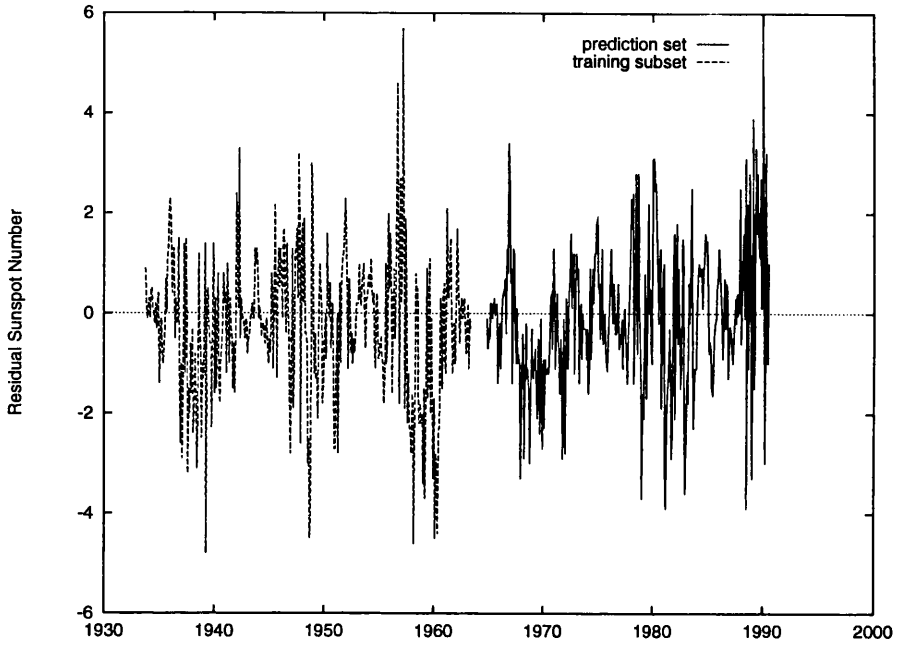
Figure 4.8: Predicted v Actual smoothed monthly sunspot number for the four cases of 1, 6, 12 and 18 month ahead predictions. In figure (a), the RMS prediction accuracy of 1.42 corresponds to predicted values which are almost indistinguishable from the actual data on the scale shown. Hence the residual and percentage errors are considered in Figure 4.9 to provide clearer analysis of these results.

consistent between training and predicting as there are no major differing features between the two regimes. The largest errors in terms of sunspot number occur during the peaks of the cycles with the smallest errors in general at sunspot minima. The picture changes when the percentage error is plotted as in Figure 4.9(b). Now the errors are noticeable lower during the maxima with the occasional very large spike occurring during the years of low sunspot number. Since the sunspot behaviour during the peak is much more important and significant than the precise value at minimum, this behaviour is appropriate to the requirements of a sunspot prediction method. The percentage error for one month ahead is shown to be within 5% except at sunspot minima where the percentage error is much more sensitive to small errors in actual value. Furthermore the percentage error during the prediction range is consistent with that achieved during training.

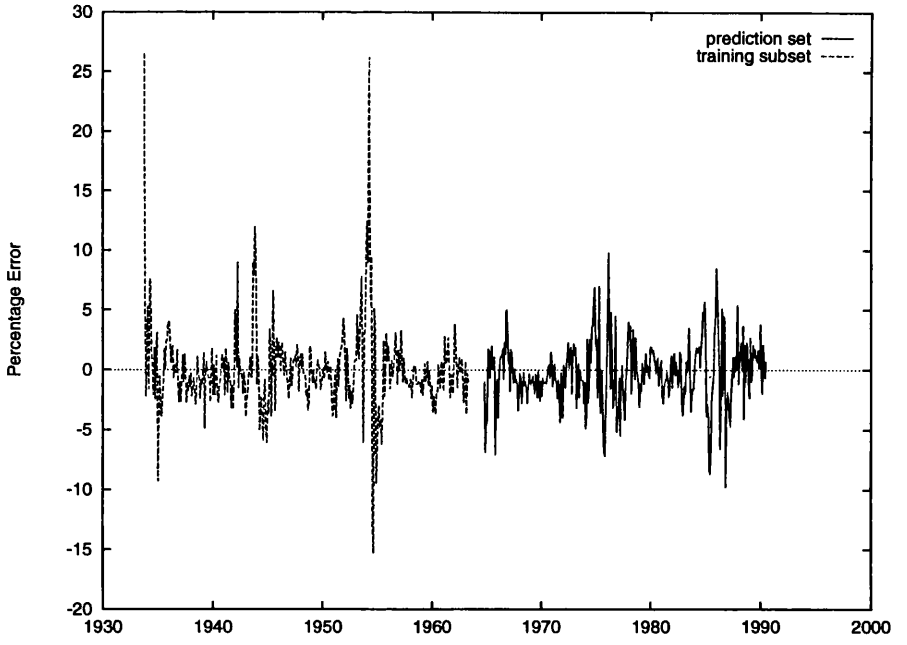
Turning now to Figure 4.10, the same two plots are now shown for the six month ahead prediction along with the graphs of percentage error for 12 and 18 months. In all cases the behaviour of the networks is consistent through the training and prediction sets. For the six month ahead predictions, the percentage error is within 15% during most of the training set and 10% during the times of solar maximum. The prediction accuracy is slightly worse in general but falls within similar boundaries. For twelve month ahead results the error bounds are about 20% except at the minima where, since the average value of the sunspot number is around 10, it is not uncommon for the prediction error to be out by this amount resulting in very large percentage errors. As mentioned before, this is not important as small differences (in terms of sunspot number) such as these at times of minima do not cause any of the problems which were discussed in Section 1.5. A similar picture occurs in the case of 18 month prediction where the percentage error bounds are about 25% during the important parts of the cycles.

4.5.2 Summary of Neural Network Analysis

This chapter has displayed through examples many of the theoretical considerations introduced in the previous two chapters where the idea of neural network computation was described. In particular, Section 2.5 raised several questions regarding artificial neural networks and this conclusions section is used to consider possible answers which the results of this chapter may have provided. Dealing with these questions in order, determination of the best architecture, including number of units in each layer, has obviously been a major part of this chapter. Some of the aspects of architecture were 'solved' through previous theoretical or experimental work. For example, only feed-forward networks with one hidden layer have been considered so far, based on the conclusions of Cybenko (1989) that such a system is capable of approximating any given continuous function. No variation of activation function was investigated and only incremental updating used in the learning phase. These effects were assumed not to play a major role in gauging how well a neural network would learn and represent the sunspot number, and instead affect only the internal

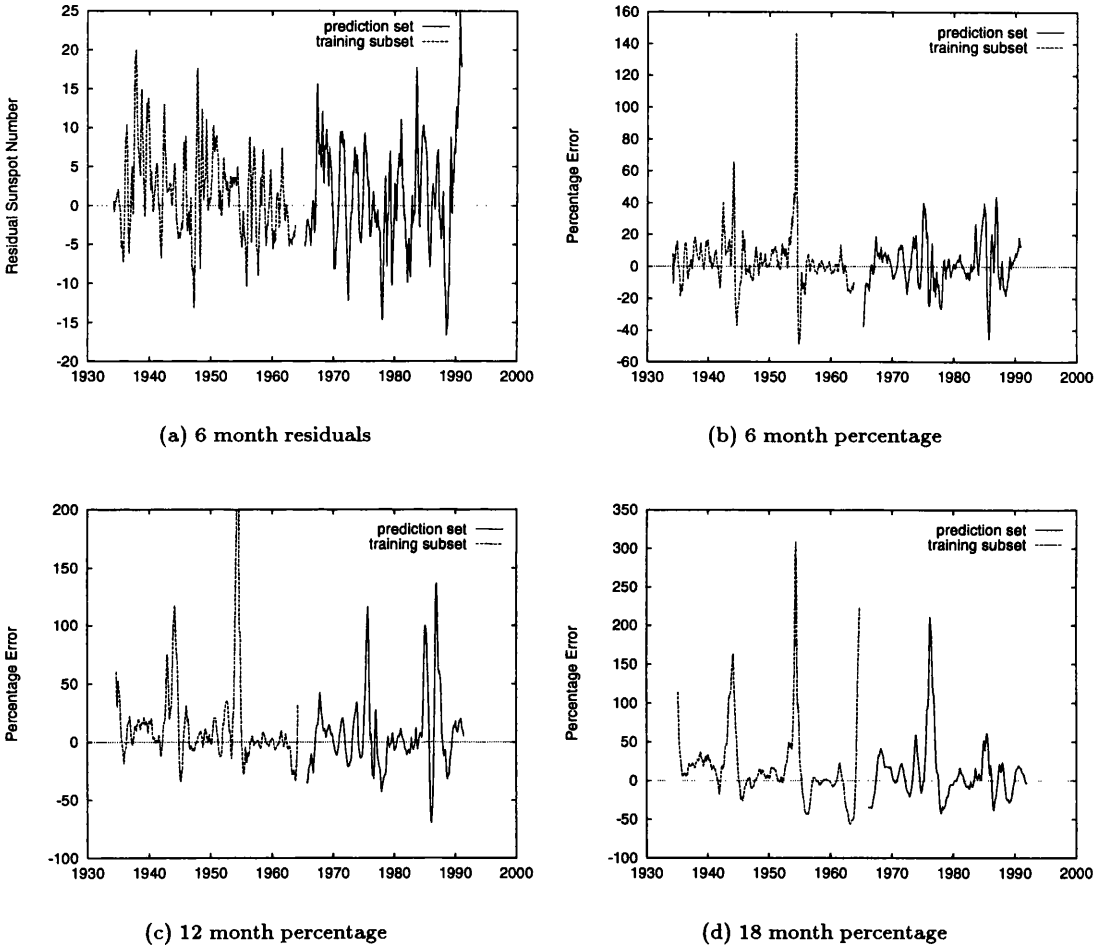


(a) 1 month residuals



(b) 1 month percentage

Figure 4.9: Figure (a) plots the residual errors for the last part of the training set and the subsequent prediction set for the 1 month ahead predictions of the 12-18-3 network. The accuracy of the results is comparable between the two sets. This provides further evidence that the network is indeed capable of generalising outwith the training data. Figure (b) shows the same features except that the percentage error is plotted.



(a) 6 month residuals

(b) 6 month percentage

(c) 12 month percentage

(d) 18 month percentage

Figure 4.10: The residual and percentage errors for the 36-27-6 network predicting 6 months ahead are plotted in Figures (a) and (b). The percentage errors on the 12 and 18 month ahead predictions are included in Figures (c) and (d).

dynamics of the neural network. The number of hidden units (n_{hid}) required has always been a major question in neural network research, and in this chapter, experiments were carried out with a wide variation in this number. From the results in Tables 4.1 and 4.5 it is not straightforward to draw any conclusions regarding a relationship between n_{hid} and the size of the input and output layers. The obvious result that too few hidden units severely restrict the learning of a problem was obtained. Once this 'minimum' number was overcome, a great variation was found for the optimal networks for different timescales of prediction. It was expected beforehand that in general a guideline for n_{hid} would be given by the bounds of the number of input and output units. This is based on relaxing the rule of thumb that $n_{hid} = (n_{in} + n_{out})/2$ is a good initial estimate. However, several examples were found which fell outside this initial estimate, most notably the 12-18-3 and 36-48-18 networks for 1 and 18 month ahead prediction respectively. Various methods, such as Optimal Brain Damage, were mentioned earlier (Section 3.5.1) for reducing the number of hidden units to obtain a network with similar generalisation abilities, or for penalising large values of n_{hid} as part of the cost function. The implementation of such techniques to the results of this chapter remains an area for research (although Weigend et al. (1992) made an initial start).

Moving onto the second block of questions raised in Section 2.5, it is fair to say that the networks trained here have certainly learnt the task in hand. The question of the amount of information required to give good performance is answered by two factors: the length of input pattern and the size of the training set. It was assumed that as long as the training set spanned a much longer time than the timescale of variability, then the data could be presented piecemeal in a large number of input patterns, whose individual length was closer to the timescale of predictions being sought. Hence the longest input pattern was 36 months for monthly prediction but the total training set covered seven solar cycles. Only a cursory examination with networks with much longer input patterns (eg 11 years or 131 inputs) was attempted and appeared to justify this assumption (although the results have not been quoted here). In any event, to attempt the same investigations with such large networks would be prohibitive in terms of learning time. The extension of the training set to cover almost the full past history of sunspot data was not found to improve the accuracy for the monthly predictions carried out here. The conclusion from this is that sufficient information about the monthly variations was included in the restricted 7 cycle span and the use of further data resulted simply in longer and more involved learning phases.

This leads onto the question of training times. In general these are different in terms of computational time for different sizes of networks, but the observed fact that the larger networks tend to reach their best generalisation point earlier partly compensates for this. The training phase is sufficiently long, however, that it requires to be carried out independently before a real time application could be considered. The possibility of retraining a network on more recent data in real time is left as an open question to be discussed. The learning rate is another parameter of

neural networks which affects the number of training iterations required. Different runs were carried out with a selection of values for ϵ to determine the most suitable value for each case. The most obvious extension from this is to introduce an appropriate method of adaptive parameters (changing the learning rate during training) so that the network has the most appropriate value for ϵ (and possibly the momentum factor as well) at each timestep during training.

Finally the generalisation ability of networks is summarised. The methods of internal validation and the check-set approach were introduced as ways of monitoring network training with the reasons necessitating their inclusion. Since all the networks were trained with a view to predicting the future sunspot number, the generalisation ability of them is the all-important feature. Thus, in agreement with the conclusions of Hertz (1993b), training must be monitored for generalisation ability and this measure used to stop training at the best point. This is confirmed by the results of Section 4.3, where it was shown that the training error on its own means nothing with respect to the best stopping point for training. Instead the emphasis falls on the validation error measure to gauge the network's generalisation ability.

Chapter 5

Prediction of Other Indices and on Other Timescales

5.1 Introduction

The previous chapter concentrated on introducing and discussing the method and techniques involved in using artificial neural networks to analyse and predict the behaviour of time series data. These ideas were put into practice using the smoothed monthly sunspot number as the test solar activity index. In this chapter we utilise the ideas already established to analyse alternative solar activity indices over different timescales. First of all, the raw unsmoothed monthly sunspot number is considered. This time series is obviously noisier and so provides a test of a neural network's ability to learn rapidly varying data. The predictions also operate on a different effective timescale than the smoothed monthly data because of the different scale of variability between the two time series.

In Section 1.4.1, the solar 10.7cm flux was introduced and its importance for solar-terrestrial interactions discussed. Hence it provides an alternative index which must be studied to provide a better understanding of solar activity levels and their consequences. Since longer term fluctuations closer to the length of the solar cycle also need to be predicted, an analysis of the time series of yearly values of the solar flux index is also completed. Since the amount of direct solar flux measurements is limited to those made from 1947, ways of expanding this database are necessary to provide sufficient data for the neural network to be trained and tested.

Finally the focus of attention is moved from the Sun to the Earth where the interaction of the solar wind with the Earth's magnetic field leads to variations in the latter and atmospheric effects such as geomagnetic storms and aurora. Geomagnetic activity is discussed briefly along with the indices which are used to quantify the variations. Since this data is also in a form suitable for time

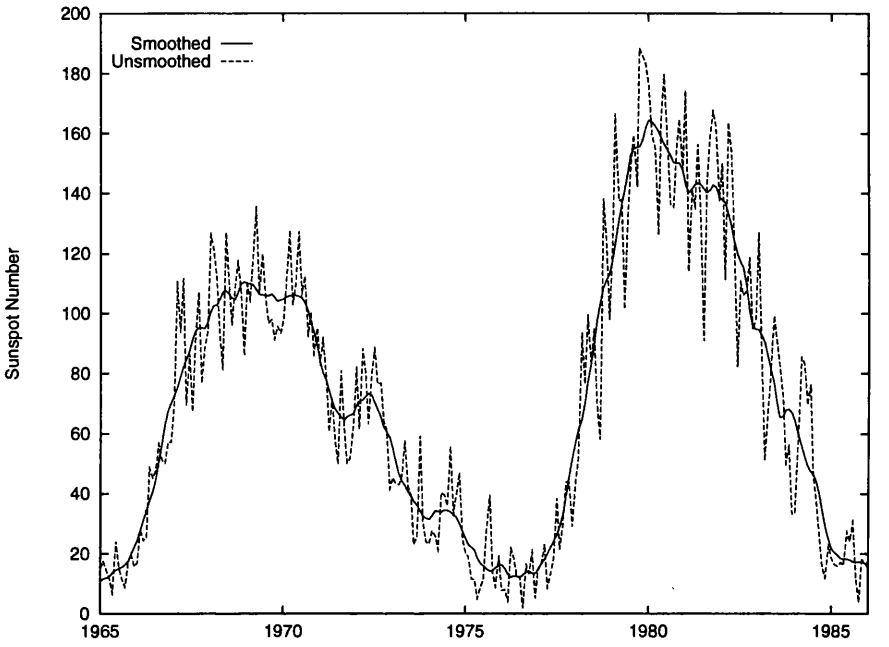


Figure 5.1: The difference in complexity between the variations in the smoothed and unsmoothed monthly sunspot number. The period from 1965 to 1986 corresponds to solar cycles 20 and 21.

series analysis, we carry out neural network predictions of one of these indices along with a brief discussion of alternative prediction methods.

5.2 Unsmoothed monthly sunspot numbers

Following the in depth study of the analysis and prediction of the behaviour of the smoothed monthly sunspot number in the previous chapter, the complexity of the underlying dynamical system is now increased as the fluctuations in the raw unsmoothed monthly number is considered. This can be seen clearly by plotting the unsmoothed and the smoothed behaviour together. Figure 5.1 shows solar cycles 20 and 21 in both smoothed and unsmoothed monthly data.

For prediction purposes, the same approach as for the monthly smoothed data is followed. The numbers of input and output units were varied between the same limits as before to make up the selection of networks to be trained and tested. From Figure 4.1, it is clear that with only a bare minimum of one unit in the hidden layer, this type of network is unable to learn in the smoothed case and so architectures with limited hidden units were not trained on the unsmoothed data. This was a slight saving on processor time. The other factors used in the initial training section (4.1.1) were kept the same for the unsmoothed data. A learning rate of $\varepsilon = 0.005$ proved to be a good initial estimate for gauging prediction ability with the possibility of improvement

Time ahead (months)	Network	Iteration	RMS Error
1	18-12-6	3200	17.73
2	36-42-18	1300	19.36
3	36-42-18	1200	19.74
4	36-42-18	1400	20.12
5	36-36-18	3300	20.49
6	36-24-18	3400	20.99
7	36-24-12	1500	21.16
8	36-24-12	1600	21.06
9	36-27-18	900	21.34
10	36-27-18	800	22.04
11	36-12-18	2400	22.55
12	36-27-18	800	23.30
13	36-21-18	800	24.10
14	36-27-18	600	24.05
15	36-12-18	1200	24.01
16	36-21-18	800	24.38
17	36-27-18	700	24.77
18	36-48-18	1000	25.28

Table 5.1: Neural network prediction results for unsmoothed monthly sunspot numbers, 1-18 months ahead. The columns show the month ahead, the network and its number of training iterations and finally the RMS error on the prediction set made up from cycles 20, 21 and 22.

over selected networks. The limited training set also appeared to provide sufficient information to allow a good prediction accuracy over the timescales of monthly prediction, without the need to train the networks on the full data range.

The results obtained are now presented in a similar format as in Chapter 4. The initial neural network results are analysed over the prediction period of cycles 20, 21 and the start of 22. These results are then compared with the McNish/Lincoln method. Contamination of the results due to a delay effect is also considered here.

5.2.1 Results

In this case the Table of results 5.1 shows the additional complexity in predicting the unsmoothed sunspot numbers. The accuracy achieved is considerably less than the results achieved in the previous chapter for smoothed monthly sunspot prediction. The initial impression from these results is that the complicated nature of the rapid variations in the unsmoothed data require the networks to be larger and thus have more connections which can vary to fit the training data. It was also found from sorting the best results that the best networks were far less well defined for this data, with many networks producing a very similar prediction accuracy. The reason for this will be due to the nature of the data. The noisy nature of the time series will place more of a limit on the predictability which can be attained for it. This limit may be reached by the networks before the differences in architecture become significant in making one network better than another. As

an example of this, simply for one month ahead predictions, although the best is quoted in the table as 17.73, there were 350 other results (different networks and/or different training iterations) all with an RMS error over the prediction range under 18. The benefit of these results is that one network could be found and used for the whole range of predictions from 1 to 18 months ahead as the differences from the specific optimal accuracy will be slight in most cases. An example of this is included in Figure 5.2 in the following section when the results are compared with the ML method.

Following on from the results obtained through the optimisation sections of the previous chapter, based on the conclusions from that work, no further optimisation is attempted here. This decision is based on the fact that all of the networks showed a minimum in learning within the 6000 training iterations, none of the larger networks (36 input units) had shown any quicker learning with a change in learning rate and using a full dataset gave no improvement while being computationally more expensive.

5.2.2 Comparison with McNish/Lincoln

Following on from the pattern of analysis laid out in the previous chapter, we now compare the prediction accuracy of the neural networks with that already achievable using the McNish and Lincoln method. From the nature of the ML method in computing a mean cycle to base its predictions from, it is expected that the more complex nature of the unsmoothed data will not prove well suited for this method. The prediction accuracy of the ML method is again plotted against the best networks for each month ahead prediction. In Figure 5.2, several pieces of information are shown. Initially there is the line of best network predictions for each individual month and this can be compared with the equivalent ML predictions. This would show the neural nets again to be superior to the ML method for predicting sunspot numbers. As a comparison to demonstrate the difference from the smoothed sunspot number results, this graph includes a plot from the fourth column of Table 4.2. Also included is the prediction accuracy for each of the 18 output units from the 36-27-18 network (after 900 iterations) to demonstrate the fact that this network is one of several of the 18 output nets which could be used to give predictions for all these outputs without departing too much from the optimal line.

5.2.3 Delay in Unsmoothed Predictions

The discussion of the results in the previous section is based on the assumption that they have not been contaminated in any way by the delay effect. This effect was described and tested for in Sections 3.5.3 and 4.4 respectively. The same procedure is now carried out here. We concentrate on the results from the 36-27-18 as these follow the same general behaviour of prediction accuracy as the optimal results from various networks.

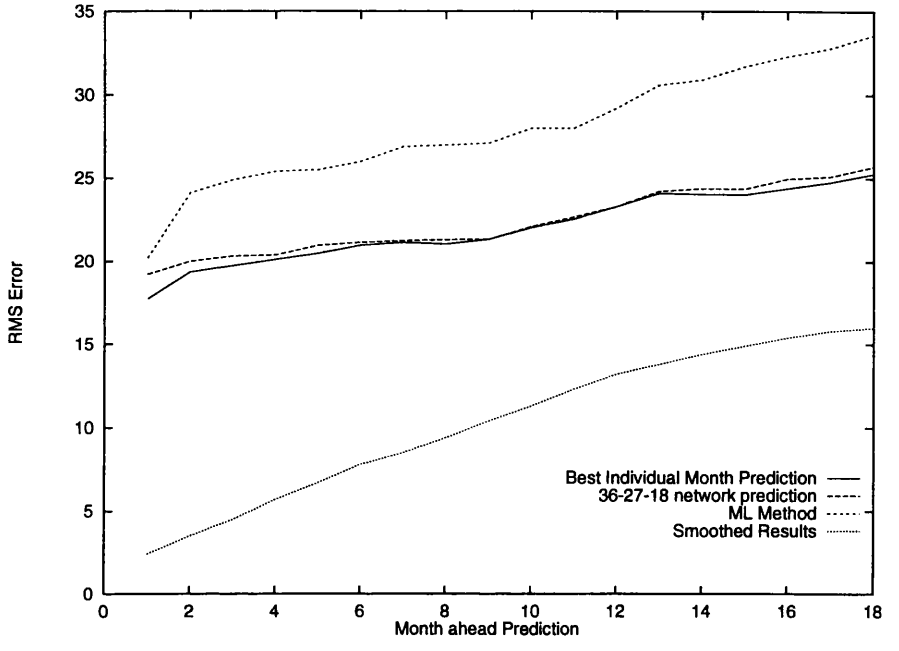


Figure 5.2: The best neural network predictions for the unsmoothed monthly sunspot numbers. Also plotted for comparison are the McNish and Lincoln prediction accuracy, the smoothed sunspot number results and the accuracy from each output of the 36-27-18 network.

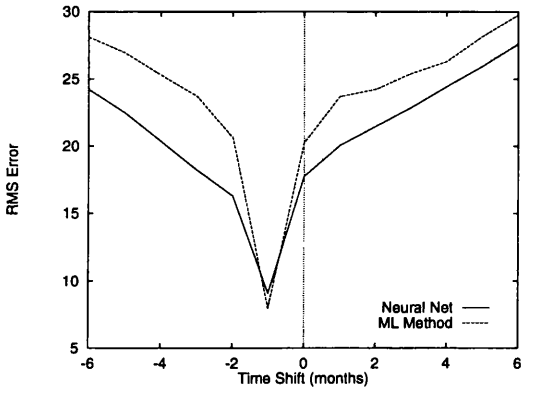
Immediately from Figures 5.3 it is obvious that there is more of a problem with neural network prediction using this dataset. The sharp minimum in the first two figures (5.3(a) and 5.3(b)) corresponds to a time shift of exactly the same as the prediction time in each case. This might suggest that the network is simply echoing the inputs it received as the best way to minimise the error on the training set. For the 6, 12 and 18 month ahead predictions, there is not the obvious sign of direct echoing although the small dips at -6, -12 and -18 on the x-axis of Figures 5.3(c), 5.3(d) and 5.3(e) indicates that a contamination of the prediction ability is certainly present. Also plotted in these figures are the equivalent results for the McNish and Linclon method. These show an even more considerable drop off for the precise prediction period ahead with an RMS figure less than the network for the first three graphs.

Looking at this in more detail by considering sections of the actual predicted values, evidence of the delay shows up quite clearly. For example, quoting the actual values and the neural network predictions for one month ahead shortly after the onset of cycle 20 shows up the following behaviour:

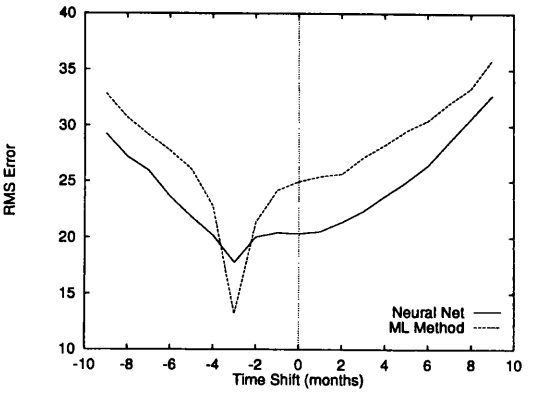
Actual 57.2, 57.2, 70.4, 110.9, 93.6, 111.8, 69.5, 86.5, 67.3, 91.5, 107.2, 76.8

NN Pred. 56.0, 58.7, 61.9, 74.9, 103.6, 97.4, 115.2, 88.2, 93.4, 80.5, 94.4, 107.9

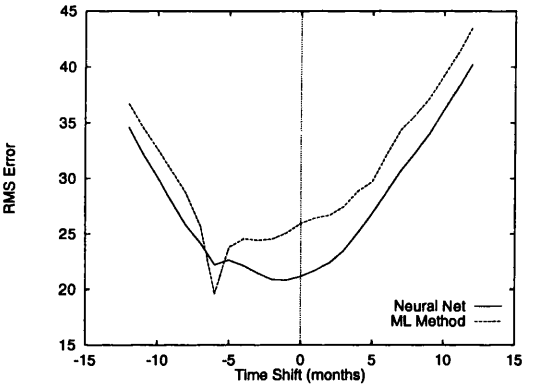
This is only a small section of the typical appearance. The trend is clearly seen in a plot of Actual v Predicted values (Figure 5.4(a)), concentrated on the maximum phase of solar cycle 21. The



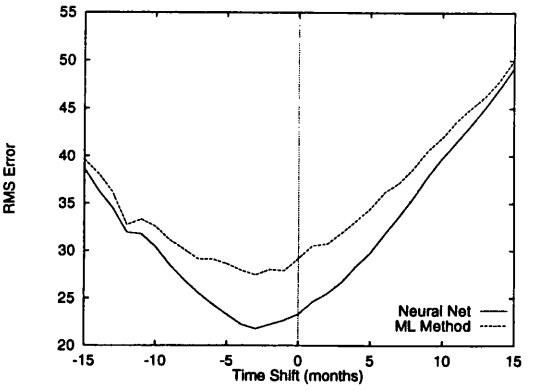
(a) 1 month ahead



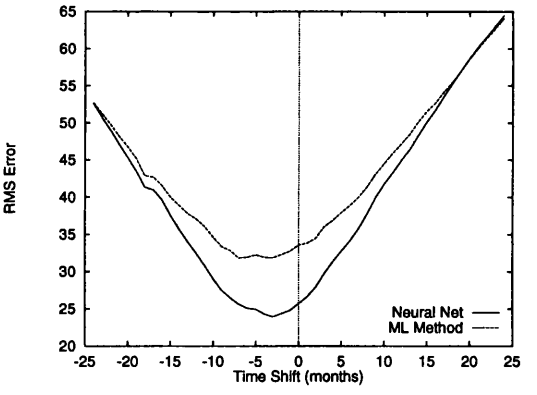
(b) 3 months ahead



(c) 6 months ahead



(d) 12 months ahead



(e) 18 months ahead

Figure 5.3: The results of time-shifting analysis carried out on the unsmoothed predictions 1,3, 6, 12 and 18 months ahead. The first two figures suffer from an obvious delay equivalent to the prediction time, while the other three show features in the plot at the prediction time.

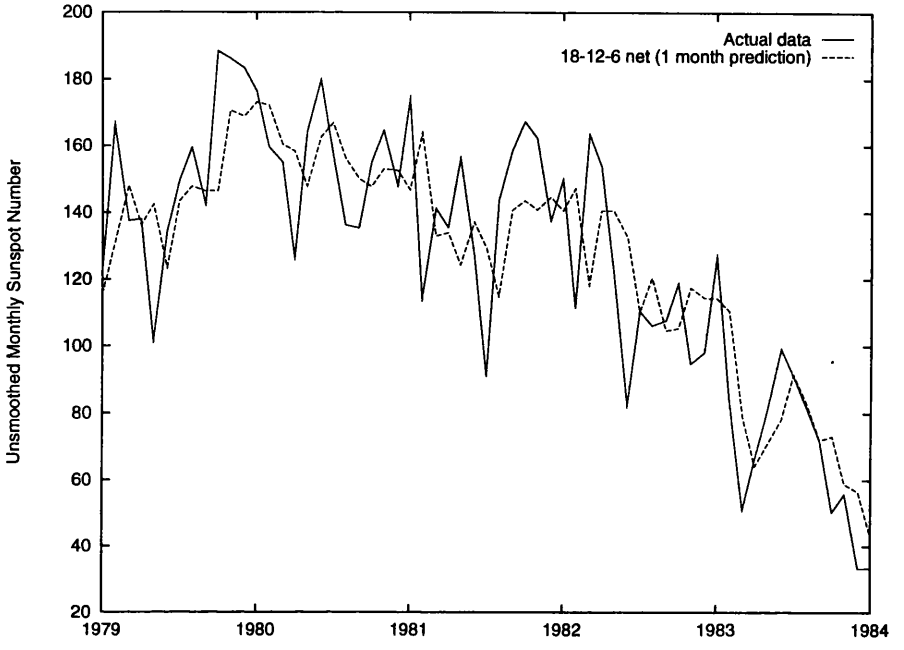
predicted values are delayed by one month and also the network learning has partly smoothed out some of the rapid variations in the month to month behaviour. Since there appears to be some possibility of echoing involved in these results the section is concluded by considering the prediction accuracy which could be achieved through simply taking the last input as the predicted value for the next point in the time series, ie echoing the inputs. The form of Figure 5.2 is reproduced in Figure 5.4(b) where the best network predictions and the ML method's accuracy are compared with that obtained through an echoing procedure. This method represents essentially the simplest prediction scheme possible. In addition, the RMS Error was also calculated for all 1-18 month prediction with the network timeshifted backwards for all these values and plotted in Figure 5.4(b). This represents the accuracy that could be achieved if the networks were corrected for the delay. For echoing to be present in the predictions, these RMS errors must be close to those produced through the echoing procedure. The conclusion to be reached from the various plots is that the predicted values are definitely *delayed* up until four months and so a correction method would drastically improve the accuracy which is achievable. Thereafter the network predictions provide the best representation of the future behaviour of the time series. The effectiveness of the echoing method is shown to be inferior at all points, including the first few months where it had been suggested to be occurring. This demonstrates in practice the difference between delayed and echoed predictions. The importance of this difference is that if echoing is taking place then the network output is worthless while if a correction for the delay *can* be made then the predicted values will prove much more useful. A more extreme example of this difference will be found later when geomagnetic activity data is considered in Section 5.4.

5.3 Prediction of Solar Flux Activity

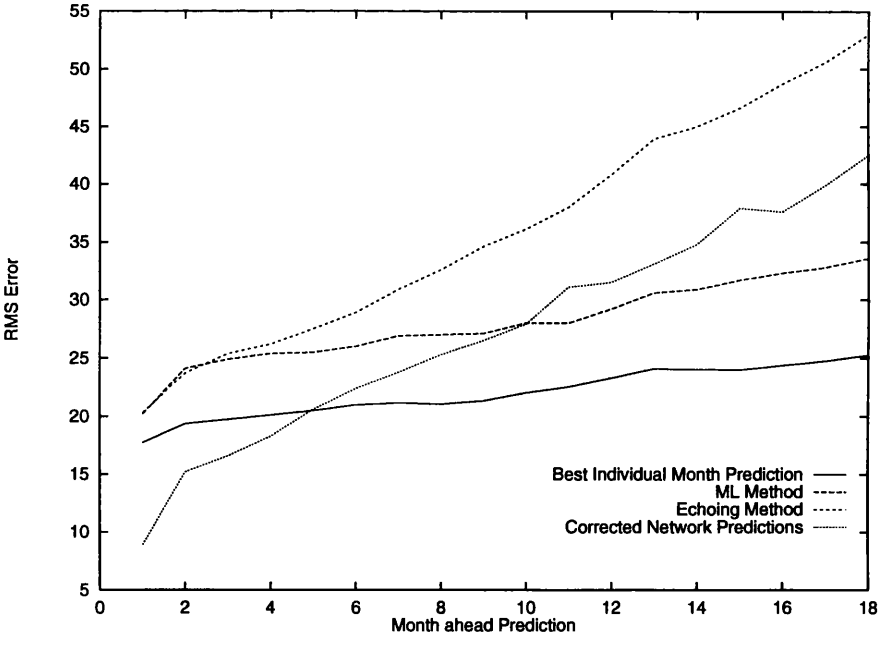
The focus now moves away from consideration of the sunspot number as a measure for describing variations in solar activity. While the variation in sunspot number is useful for, eg space weather forecasts, through its correlation with other solar cycle effects, the measurement of the solar radio flux is a more direct method of predicting solar activity and its effect on the terrestrial environment. These effects were discussed in Section 1.4.1.

5.3.1 Smoothed Monthly Numbers

The problem of predicting smoothed solar 10.7cm flux variations is considered in two ways. The correlation between this index and the smoothed sunspot number over the period 1947 to the present day has already been described (see Section 1.4.1 and Equation 1.4). For variations on the scale of 13-month smoothed measurements, the measurements of the solar flux since 1947 do not provide sufficient data for a large enough training set to be established (see Section 4.1.1 on



(a) 1 month predictions v actual data



(b) Summary of prediction accuracies

Figure 5.4: Figure (a) plots a section of the 1 month predictions against the actual data. A delay in prediction of many of the features is clearly visible. Figure (b) summarises the prediction accuracies using the various methods described in the text.

choosing a suitable size of training set for sunspot number analysis) while still allowing the same prediction range over cycles 20-22. The result of Section 4.2.5 showing that the restricted 7-cycle training set provided enough information for prediction does, however, have implications for solar flux prediction. Since over four cycles of data is available for this index, it may be possible to use this data as the training set when predictions beyond the current time are required. For now, while it is still necessary to test the prediction accuracies achievable using neural networks, ways of extrapolating the data are considered. This is done in the knowledge that the prediction accuracy might be slightly different than when only real measurements are used. The strong correlation with the sunspot number suggests two possible ways of extending this data to provide a larger time series history. The quick method for doing this is simply to use Equation 1.4 to translate the predicted sunspot numbers already obtained into predicted values for the solar flux. A more thorough approach is to extrapolate the solar flux values back in time assuming the correlation with the sunspot number to generate a past history of solar flux values complete from solar cycle 1. Networks can then be trained on this data and used to make predictions in the established manner.

Solar Flux Prediction 1

Consider first the case of predicting the smoothed monthly sunspot number and then converting this into a solar flux prediction. For the specific cases of 1, 6, 12 and 18 months ahead quoted in the previous chapter, the sunspot number generated for the prediction interval over solar cycles 20, 21 and the start of 22 is taken using the appropriate best network. These values are translated into estimates for the solar flux and compared with the actual measurements over the test range. The results are shown in Table 5.2(a).

Solar Flux Prediction 2

Since the variations in solar flux show the same sort of variations as the sunspot number, only the networks which were used in the optimisation sections of the previous chapter were trained on the extrapolated solar flux data. It was assumed that these architectures would also have shown up amongst the best if all the network combinations of input, hidden and output units had been trained. Also, based on the conclusions from the section on using the full history of data as a training set (Section 4.2.5) that for predictions on these timescales, it is not necessary to include all the data, the approach of training on cycles 13-19 and using cycles 5-11 as a test-set was continued. The best results are presented in Table 5.2(b). Looking at both tables to compare the different approaches, it is seen that on the first prediction timescale of one month ahead there is little to choose between methods. Thereafter the more thorough approach of training the networks on the specific solar flux data appears to pay off better.

Network	Prediction Ahead	RMS	χ^2
12-18-3	1 month	1.30	3.64
36-27-6	6 months	6.39	93.1
36-21-18	12 months	12.09	337.2
36-48-18	18 months	14.67	472.5

(a) Sunspot to solar flux prediction

Network	Prediction Ahead	Iteration	RMS	χ^2
12-18-3	1 month	43300	1.33	3.67
36-27-6	6 months	44200	6.21	92.3
36-48-18	12 months	5000	11.27	299.8
36-48-18	18 months	3700	13.97	454.8

(b) Direct solar flux prediction

Table 5.2: Tables (a) and (b) represent the predictive accuracy of the methods of solar flux prediction 1 and 2 respectively. Apart from the 1 month ahead predictions which are very close, the accuracy is improved when the networks are trained extensively on an extrapolated solar flux dataset.

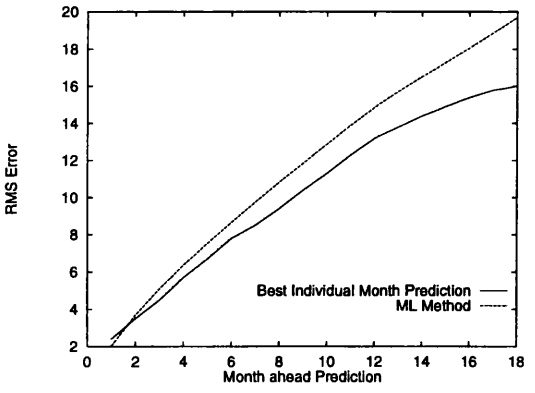
Comparison with McNish/Lincoln

To complete this section on monthly solar flux predictions we compare the prediction accuracy of the neural networks against the chosen standard benchmark of the McNish & Lincoln method. Also included are plots showing actual v predicted values. Table 5.3 shows the compared results for the neural net predictions against the ML method. These results are equivalent to Figure 4.3(a) for monthly sunspot number prediction. This figure is reproduced here along with a graphical form of Table 5.3 to allow for comparison. Since the correlation between the sunspot number and the solar flux forms an integral part of obtaining these results, it is not surprising that the general form of Figures 5.3.1 and 5.3.1 is the same. This does give some further backing to the consistency of the method of neural networks for these predictions. The prediction accuracy over the solar flux values is better than the equivalent values for the sunspot numbers. This is consistent with the results of Mugellesi and Kerridge (1991) when they compared their predictions using the variant of the ML method also employed here. One suggestion for this is the fact that the solar flux values are inherently smoother than the sunspot number due to the more quantitative nature of this index. This is confirmed by taking correlations of the smoothed and unsmoothed numbers for each dataset. Over the period in which both indices were measured, that is 1947-present, the RMS deviation on the smoothed and unsmoothed solar flux is 23.1 while 25.4 is the equivalent measure for the sunspot number.

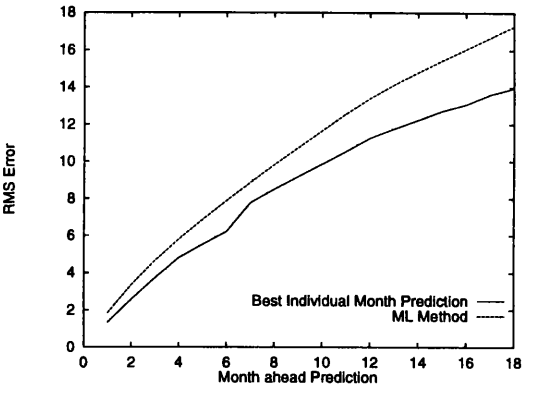
This section is completed with a selection of graphs showing the predicted solar flux values versus

Time Ahead (months)	Best Net RMS	ML RMS
1	1.33	1.84
2	2.58	3.36
3	3.73	4.66
4	4.81	5.80
5	5.52	6.86
6	6.21	7.88
7	7.77	8.87
8	8.50	9.83
9	9.19	10.75
10	9.86	11.66
11	10.54	12.57
12	11.27	13.41
13	11.77	14.15
14	12.24	14.81
15	12.73	15.44
16	13.08	16.05
17	13.62	16.66
18	13.97	17.28

Table 5.3: Comparison of the best neural network predictions when trained on the extrapolated solar flux against the accuracy achieved by the McNish and Lincoln method. The neural nets are better in all month ahead predictions.

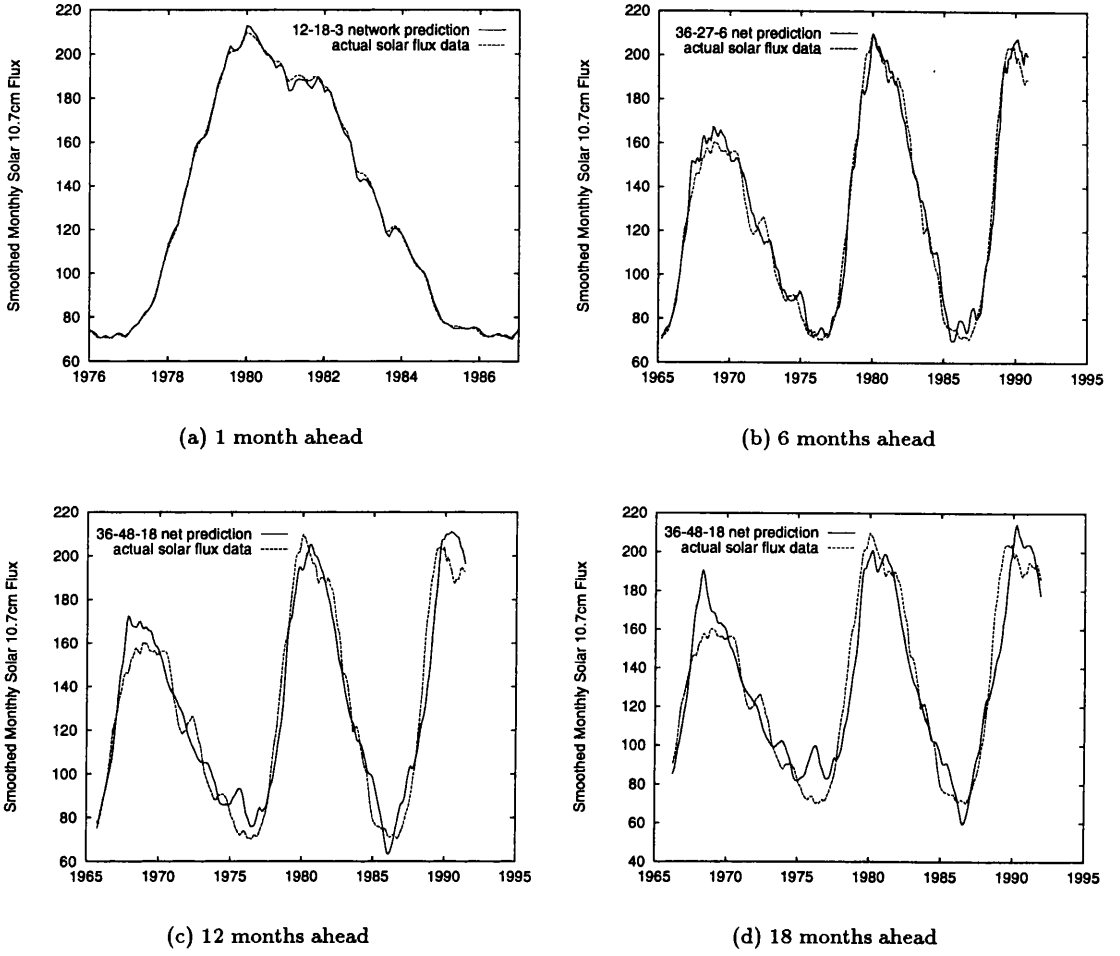


(a) Neural net v ML for sunspots



(b) Neural net v ML for solar flux

Figure 5.5: The comparison plots for neural networks versus McNish and Lincoln for smoothed monthly sunspot number and smoothed monthly solar flux values. Since these indices are well correlated, it is not surprising that similar results are achieved. This does give some reassurance on the consistency of neural network predictions.



(a) 1 month ahead

(b) 6 months ahead

(c) 12 months ahead

(d) 18 months ahead

Figure 5.6: These graphs show plots of the predicted versus the actual solar flux for the prediction range of cycles 20, 21 and 22. Figure (a) is restricted in range to cycle 21 in an attempt to make the differences in the graph discernable.

the actual data for particular month ahead predictions. These results are taken from those obtained through the more thorough approach of training the networks on the extrapolated solar flux, rather than from a conversion of predicted sunspot numbers. Figure 5.6 selects the 1, 6, 12 and 18 month ahead cases as examples. In the first figure (5.6(a)), the prediction interval is restricted to solar cycle 21 in an attempt to make the actual and predicted lines discernable on the given scale.

5.3.2 Yearly Predictions

So far the analysis has centred around the medium length timescale of predictions on the order of months, although the prediction of smoothed and unsmoothed sunspot numbers does include a difference in effective prediction timescale. In this section, however, attention is focussed on the longer term behaviour of solar activity on a par with the variation of the solar cycle itself. The

average yearly solar flux value is plotted against the yearly sunspot number over the period from 1947 to recent measurements in Figure 5.7(a). The best fit least squares line is given by

$$F_y = 0.915 * R_y + 59.08 \tag{5.1}$$

where this time F_y is the yearly solar 10.7cm flux and R_y the yearly sunspot number. This relationship, although it does not differ much from Eq 1.4, is used to extrapolate the flux measurements to provide a yearly measure for the full range of solar cycles back to 1755. Also shown in Figure 5.7(b) is the yearly measurements plotted against the unsmoothed monthly values of the solar flux. Due to the sparse number of points the yearly 'curve' shows up with sharp corners although it is still obviously smoother than the raw monthly data.

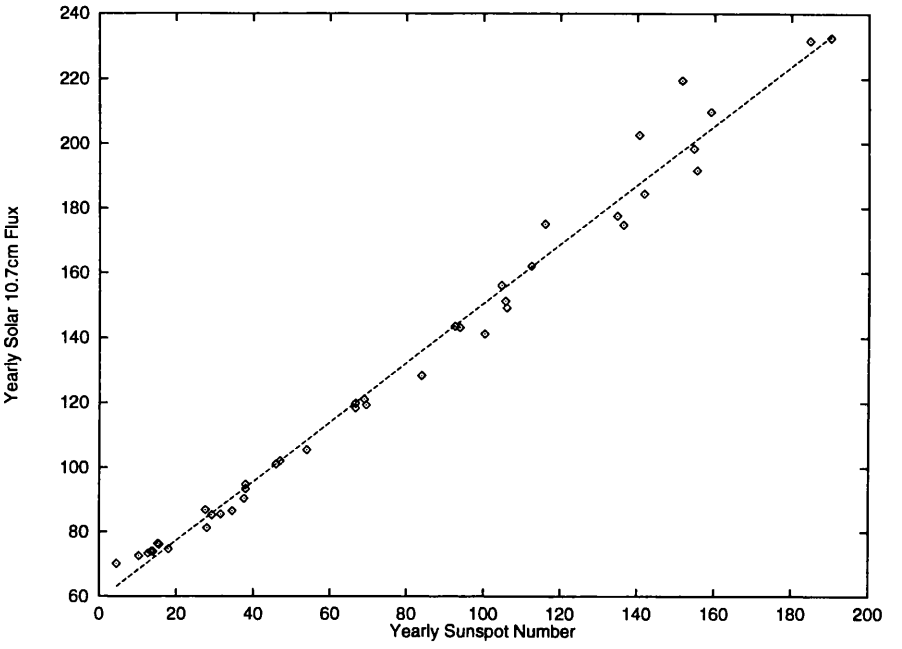
In this section, we shall discuss the prediction of yearly solar 10.7cm flux values in a similar manner to the procedure established previously for monthly (smoothed or unsmoothed) predictions. Neural network analysis of yearly values of the sunspot number have been carried out by other authors recently (see Weigend et al., 1992 and Calvo et al., 1993). Comparing actual results is obviously meaningless as different indices were used in the studies. However the other works contained significant results regarding the type of approach used to predict more than one month ahead and in this case the general behaviour of results can be compared. Since this involves a discussion of iterated single- and multi-step prediction, it shall be left until the appropriate section in the next chapter (see 6.2). Here the standard analysis including a comparison with McNish & Lincoln predictive ability will be presented.

Training

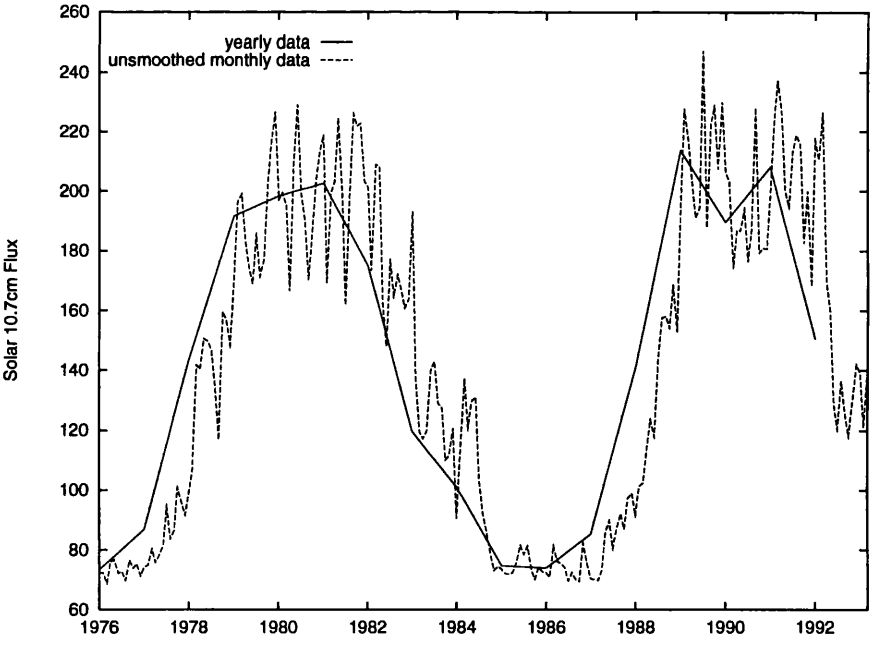
In proceeding with the training of networks on yearly data, initially the number of input units was set to 6, 12 or 18, corresponding to roughly a half, a full and one and a half cycles as the input pattern. The number of output units was set to include 1, 3, 6 and 12 units. As before, the initial study chose a wide range of hidden units for each combination of input and output layers. The range of training data which provided the training patterns was set to span cycles 3 to 19 with cycle 1 being utilised as a validation set if necessary. The prediction range was over cyles 20 to 22, the same as for monthly data.

Results

In quoting the results obtained from these trained networks, we start first with the standard measures of prediction accuracy which have been used so far. The long term implication of yearly predictions allows scope for further analysis with regard to the possible prediction of the next solar cycle based on the general ability of these networks to predict the shape of the test cycles, namely 20, 21 and 22. Table 5.4 gives the prediction errors for the various networks which were the best



(a) Yearly Flux v Yearly Sunspot No.



(b) Yearly v Monthly Flux

Figure 5.7: Figure (a) shows the correlation observed between the yearly solar flux and the yearly sunspot numbers. Figure (b), meanwhile, shows the difference in appearance of yearly versus monthly solar flux data.

Year ahead	Network	No. of Iterations	RMS	χ^2	ML Rms	ML χ^2
1	18-12-12	600	15.09	46.9	14.82	44.4
2	18-18-12	2100	20.83	84.4	18.09	67.6
3	18-9-12	6000	24.67	119.6	23.60	128.5
4	18-24-12	100	27.81	141.4	36.26	285.5
5	18-24-12	100	28.51	144.0	40.98	292.9
6	6-9-6	6000	28.64	192.6	43.49	303.6
7	18-12-12	6000	28.11	134.7	42.28	250.6
8	18-12-12	6000	28.57	134.0	51.14	302.0
9	18-24-12	3000	29.04	131.3	54.19	322.3
10	18-24-12	2900	29.66	133.0	57.55	339.1
11	18-24-12	3300	30.83	143.6	59.61	350.8
12	18-12-12	3000	33.39	166.0	61.49	354.4

Table 5.4: Results of yearly solar flux predictions covering one solar cycle in advance. The comparison between the neural network results and the McNish and Lincoln predictions shows the latter superior for the first 3 years prediction before the networks become dominant thereafter.

for the different year ahead prediction. In general it is seen that 18 inputs with 12 outputs proved the most successful combination with only the one exception for 6 years ahead. Even in this case the difference between the accuracy of the 6-9-6 net and the best 18-24-12 net is small compared to the total RMS error (28.64 to 29.17).

Predictions were also carried out for the yearly sunspot number using an adapted version of the code that was written to implement the McNish and Lincoln technique. The changes to deal with yearly data are mainly programming reasons and the basic method as described previously still holds. The accuracy of this method is quoted in columns 6 and 7 of the same table. The 1, 2 and 3 year ahead predictions are superior to the neural networks but beyond this the ML method deteriorates in accuracy considerably. In fact the predictions for longer than 8 years ahead are not much better than would be achieved by using the mean value of the solar flux over the three cycles tested as the predicted value.

It is possible to compare these results with those achieved using the monthly data for the first one or two predictions. Using monthly data to predict one year ahead (the 12 ahead prediction from Table 5.3) gives an RMS error of 11.3 compared to 15.1, considerably better. This is perfectly understandable based on the fact that the networks trained on monthly data have had much more training examples showing the typical variation of the solar flux over a year. The extrapolation of the monthly prediction accuracy to 24 months would still look to be better than the 2 year ahead prediction based on the increase of the RMS error with respect to month ahead in column 2 of Table 5.3. Any attempt to anticipate beyond that point would be meaningless and in any event to carry out a direct prediction using monthly data would require a network with an impracticable amount of output units. Instead, to carry on predicting ahead with monthly data, the approach indicated resorts to iterating the network and this possibility is considered in Section 6.2.3.

Cycle	Year of Max.	Max value	1 year	3 years	6 years	12 years
20	1970	156.1	1968 - 174.0	1968 - 190.7	1967 - 208.3	1968 - 216.7
21	1981	202.6	1980 - 198.5	1981 - 167.2	1979 - 163.2	1980 - 167.1
22	1989	213.6	1990 - 209.3	1990 - 206.9	1990 - 195.4	1990 - 175.0

Table 5.5: Table of maxima predictions over cycles 20, 21 and 22.

Network	9 years ahead	10 years ahead	11 years ahead	12 years ahead
18-24-12	2000 - 168.1	2001 - 177.3	2000 - 179.2	2000 - 207.1
18-12-12	2001 - 164.1	2001 - 170.7	2000 - 177.8	2000 - 201.7

Table 5.6: Predictions for year and magnitude of solar cycle 23 maximum for yearly solar flux values.

The opportunity exists with this data to make a start on the idea of long term prediction of the next solar cycle. For the three cycles used as the prediction set, the year and magnitude of the yearly solar flux maximum is predicted at intervals from 1 to 12 years ahead. (This is a slight cheat as to use this for prediction, the number of years to maximum has to be known already, so that the correct output unit can be used. This problem is avoided when the predictions are iterated using a fixed starting point and using the network output as subsequent inputs. This is examined in the next chapter.) The results obtained by the present approach are shown in Table 5.5. The accuracy of the results is inconsistent from one cycle to another and so it is hard to draw any specific conclusions from this table. In general the predicted magnitude values are best for cycle 22, with even the 6 year ahead value being within 10% of the actual value. The time of maximum for cycle 22 was always predicted as being one year later than occurred although this is not so surprising as this cycle peaked after one of the shortest rise times on record. For cycles 20 and 21, the magnitude of maximum is fairly well predicted one year ahead but the accuracy has dropped to a 20% error for the 3 and 6 year predictions. Prediction of the year of maximum could not be considered particularly good since the results show an error of 1 or 2 years in many cases. If a mean year of maximum is calculated from all the previous cycles then the deviation for most previous cycles will not be much more than 1 or 2 years in any event, since most cycles tend to peak before the halfway point in a cycle.

Using the 11 and 12 year ahead networks, a prediction can be made for the maximum value of solar cycle 23. These values are shown in Table 5.6. Within each column, the predictions of the two networks are generally consistent with each other, the only real exception being the difference in year for the 9 ahead column. The prediction tests in Table 5.5 do not give sufficient information regarding accuracy to allow any truly confident statements to be made about these results. Since the test predictions were only within about $\pm 20\%$ and $\pm 25\%$ for the 6 and 12 year ahead predictions respectively, these are the only bounds that can be tentatively placed on the

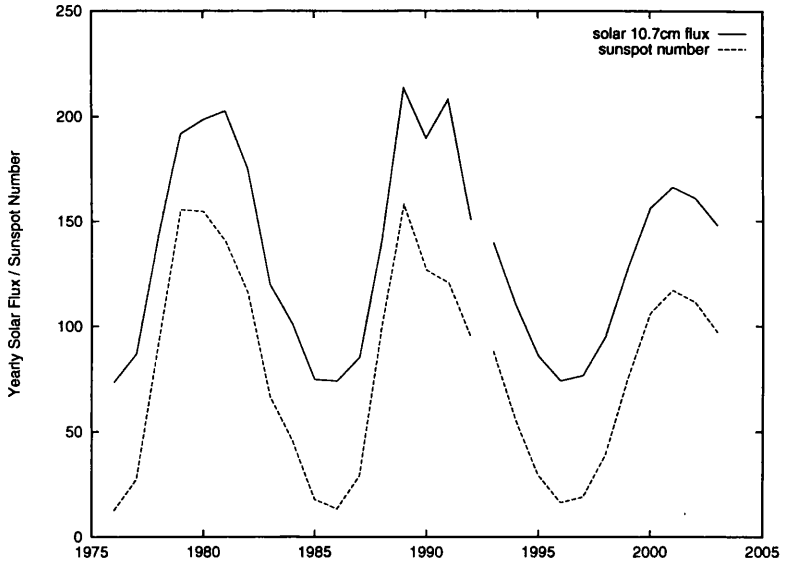


Figure 5.8: Yearly solar flux and reconstructed yearly sunspot predictions for the next 12 years, encompassing the evolution of solar cycle 23. The maxima are respectively 166.1 and 117.0 to occur in the year 20001. The actual data for cycles 21 and 22 is included for comparison.

results. The further problem is that the predictions increase considerably from the 9 to 12 year ahead estimate of the maximum. Taking the last available actual data as input to the 18-12-12 network, the twelve outputs are calculated to provide a prediction for the evolution of solar cycle 23. Figure 5.8 shows these predictions along with the yearly sunspot numbers reconstructed from these values using Equation 5.1. The predicted maximum is 166.1 to occur in 2001, giving a yearly sunspot maximum of 117.0. This would be a smaller than cycles 21 and 22 (maximum values 155.4 and 159.0 respectively) but closer in size to solar cycle 20 (maximum 105.9 in 1968). Based on the accuracy of the predictions for the three test cycles, the error on this result is estimated to be of the order of $\pm 25\%$. Since the year of maximum was not predicted to high accuracy for the test cycles, the best that can really be said about the time predictions is that they are consistent and in agreement with the general timescale of the approximate 11 year period of variation.

This covers a first attempt at estimating the long term predictive ability of neural networks. Here 'long term' is used essentially to cover the ability to predict the time and magnitude of the next solar cycle. This subject is examined in more depth in Section 6.4.

5.4 Analysis of Geomagnetic Indices

5.4.1 Geomagnetic Activity and Indices

The main index of activity is the internationally agreed K_p or 'planetary 3-hourly index' which is measured by 12 observatories situated at geomagnetic latitudes between 48°N and 63°N. After correction for latitude, these observations can be summed over a day to give the daily planetary index K_p. This index is expressed by a value of 0 to 9, with intervals of 1/3. Corresponding to these values is the 'planetary amplitude index a_p' which is interpreted as the mean amplitude variation on a linear scale (with units of 2nT). However, the values of these indices exist in time series form only since 1932.

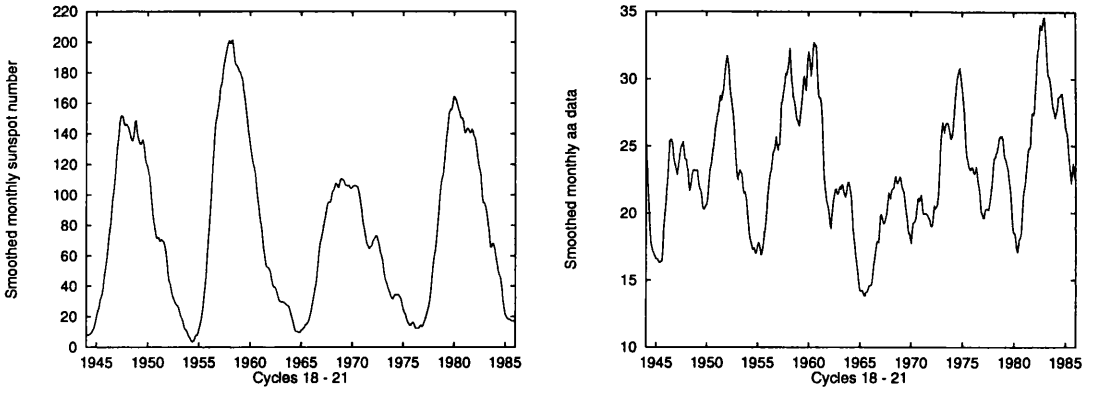
The 'antipodal index *aa*' is obtained by the observing stations of Hartland (UK) and Canberra (Australia) and has been calculated retrospectively to 1868 (Mayaud, 1975). Thus it covers approximately 11 solar cycles and so in principal can provide more information on past correlation between solar and geomagnetic activity than the more limited K_p data. Computing a least squares fit on the a_p and *aa* indices also shows a correlation with a fit of

$$a_p = -3.71 + 0.816 \quad aa \tag{5.2}$$

This latter correlation is quoted from Mugellesi and Kerridge (1991).

While sunspots themselves will have virtually no effect on geomagnetic activity, other parameters which do affect the terrestrial environment are modulated with the sunspot cycle. Thus by association, we look for similar behaviour in the geomagnetic indices as with the sunspot number. Figure 5.9 plots yearly sunspot numbers and annual values of the *aa* index over the period since 1868. This shows a clear modulation of geomagnetic behaviour with the sunspot number. However, the peaks of the *aa* do not match exactly with those of sunspot maximum, and in fact can be seen to be double-peaked in most cases. The major peak tends to occur during the declining phase of the sunspot cycle with the secondary peak occurring around the time of maximum in the yearly sunspot data. Subsequently, the minimum of this geomagnetic index lags the sunspot minimum by approximately one year (Hirshberg, 1973). One other noticeable feature of the geomagnetic behaviour is the apparent increase, particularly in minimum values, from the years 1900 to 1960. Feynman and Crooker (1978) modelled this as the same 11-year variation but superimposed on a base line value which was increasing linearly with time.

A scatter plot of the sunspot and geomagnetic data (see Figure 5.10) shows that the linear correlation is low, although a relationship of some sort definitely exists. Specifically, large values for the sunspot number seem to exclude the possibility of low levels of geomagnetic activity, although the reverse is not necessarily true. The equation of the line which can be drawn to represent the



(a) Monthly sunspot data, cycles 18-21

(b) Monthly *aa* data, cycles 18-21

Figure 5.9: Monthly smoothed sunspot and geomagnetic aa data over solar cycles 18 to 21. A modulation of geomagnetic behaviour with sunspot number clearly exists.

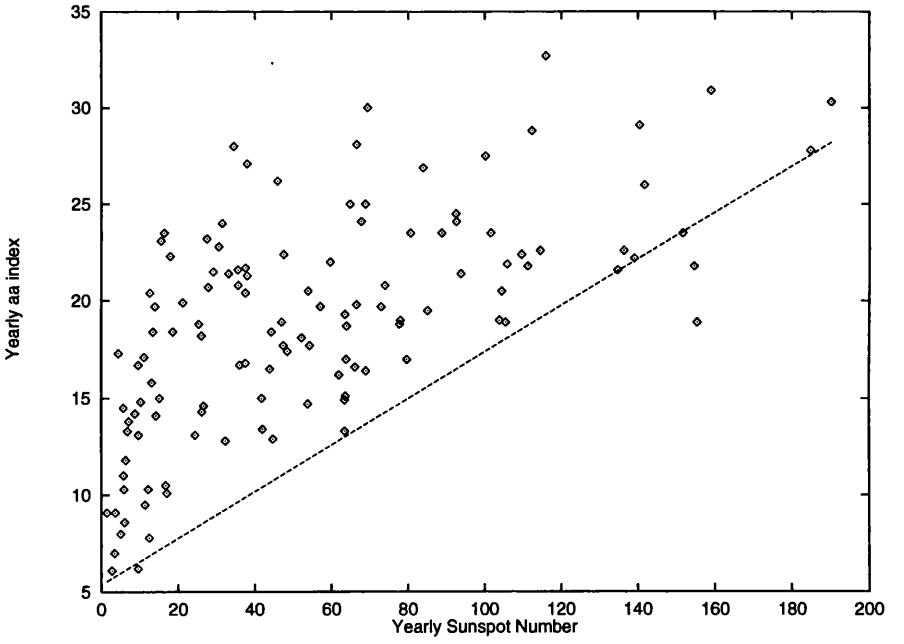


Figure 5.10: Annual averages of the geomagnetic *aa* index versus sunspot number.

minimum activity for a given value of sunspot number is the following (from Feynman, 1982):

$$aa_{(min)} = 0.12R + 5.38 \tag{5.3}$$

Gorney (1990) also states that considering only the sunspot maximum, the corresponding value of *aa* index is usually near the minimum, implying that the solar activity is less effective at coupling to the terrestrial system during solar maximum years than during the declining phase. This behaviour can be best explained by the tendency for strong solar wind streams to be generated late in the solar cycle.

5.4.2 Prediction of Geomagnetic Behaviour

Feynman and Gu (1986) proposed two methods for predicting the behaviour of the geomagnetic *aa* index over timescales of one to ten years. The first builds on the recognition of the 11-year variation superimposed on a base line that changes with time. The basic steps involved in making a prediction of a future value of the mean annual value $< aa >$ are as follows:

(i) the base value is estimated for all the values in the $< aa >$ time series and subtracted from the $< aa >$ values to generate remainders;

(ii) the mean of the remainders at previous $< aa >$ minima is computed, and this process is repeated for the 8 years before and after the minima to generate a table of mean values;

(iii) the prediction is made by adding the base value for the time of the prediction to the tabular value of the remainder for the appropriate year relative to $< aa >$ minimum.

Although this method was designed with annual values in mind it could be adapted to deal with smoothed 12 or 13 month running means as described for sunspot numbers.
The second method involved defining $< aa >$ into two parts $< aa >_r$ and $< aa >_i$ defined by

$$< aa >_r = 0.12 < R > + 5.38 \tag{5.4}$$

which is the value of $< aa >_r$ corresponding to points on the straight line in Figure 5.10 (cf Equation 5.3) and the residual values from these points given by

$$< aa >_i = < aa > - < aa >_r \tag{5.5}$$

These equations can be rewritten to give the desired $< aa >$ in terms of the other two quantities. Kerridge et al. (1989) studied this problem and considered the various merits of these two methods. The second method relies upon predicted values of the sunspot number as well as on the behaviour of the $< aa >_i$ component. Although this follows a periodic behaviour, it is by no means regular and so utilising this method to predict the annual *aa* value requires the ability to obtain accurate

values for two other unknown quantities. For this reason Kerridge et al did not proceed with this method. They did investigate further the 'mean value' method of the first approach and recognised its similarity to the McNish and Lincoln method (or variants thereof). Euler and Holland (1982) had also observed this possibility. Comparison of the two methods by Kerridge et al came down on the side of the ML technique as the most promising. This was used to predict values of the smoothed monthly *aa* value rather than annual averages. Following on from the previous analysis in this thesis with comparisons between neural network methods and the ML method, these previous results give justification to the use of the ML technique as a bench mark for prediction accuracy of geomagnetic indices.

5.4.3 Initial Analysis

The initial analysis for this data follows the same pattern as for the International Sunspot Number. We shall concentrate on smoothed monthly mean values of the antipodal *aa* index. Even so in order to carry out the network training on the span of 7 solar cycles as specified for the sunspot number prediction leaves insufficient data to implement the test-set approach for monitoring for signs of overtraining. Thus we use cross-validation as an alternative method which requires less data. We use eleven years of data covering 1876-1886 for this purpose. Also because the training set uses up all of the available data, there will be no further work in extending the training set to the full dataset. Since this index shows only an indirect coupling with the sunspot number, it was necessary to carry out a complete analysis on this dataset, with the same high level of computational work to train and test various architectures. The range of network architectures trained on this data complied with the selection criteria first explained in Section 4.1.1. As before, the most promising networks were established first over a fixed number of iterations. Thereafter, further training was carried out as necessary and a variety of learning rates were tried for the networks identified.

5.4.4 Results

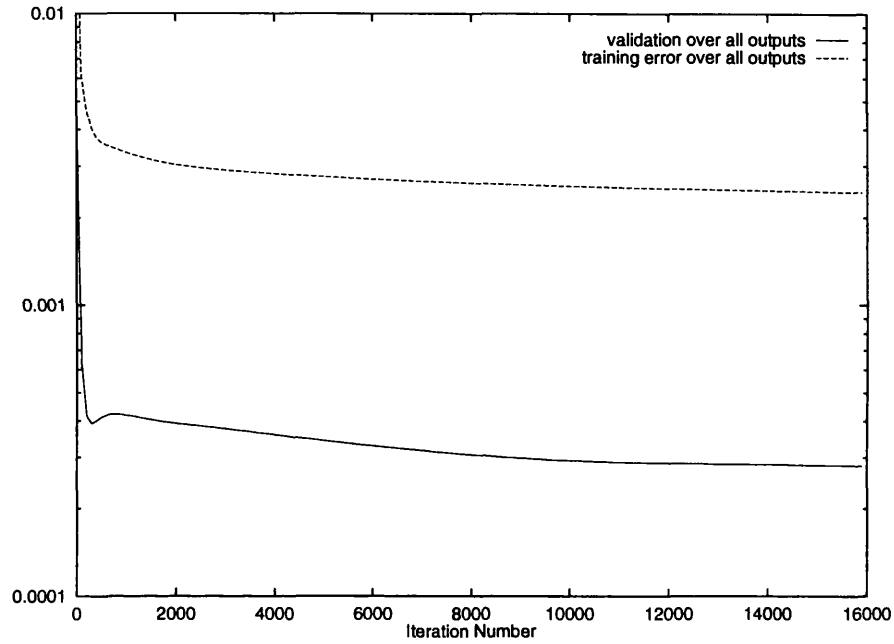
The results shown here represent the most accurate which were obtained during this study, with the different stages of network architecture selection, number of training iterations and variation of ε completed before this level of accuracy was established. Table 5.7 provides a summary of the prediction accuracy calculated for each month ahead prediction from 1 to 18 months, showing the network and the number of training iterations required before the values quoted were achieved. For all of these networks, the learning rate of $\varepsilon = 0.005$ provided these results. It should be said that the results shown in Table 5.7 were achieved through optimisation which was geared principally towards the 1, 3, 6, 12 and 18 month ahead prediction ranges. This explains the predominance of these networks as optimal for other month ahead predictions as well. This does not mean that the

Time Ahead	Network	Iteration	RMS	χ^2
1	18-6-1	15600	0.50	3.15
2	18-18-3	9900	0.93	11.40
3	18-6-3	14900	1.23	20.60
4	18-6-6	10000	1.70	39.18
5	18-6-6	4700	1.99	51.30
6	18-6-6	3900	2.31	67.21
7	36-27-18	13700	2.65	93.06
8	36-18-18	7400	2.92	113.74
9	36-18-18	7500	3.17	133.39
10	36-18-18	8000	3.39	153.08
11	36-18-18	8100	3.59	172.23
12	36-18-18	8000	3.74	187.51
13	36-27-18	6700	3.84	197.43
14	36-27-18	6900	3.90	203.90
15	36-27-18	7300	3.96	210.91
16	36-27-18	7300	4.02	218.34
17	36-27-18	9300	4.08	224.82
18	36-27-18	8400	4.16	236.30

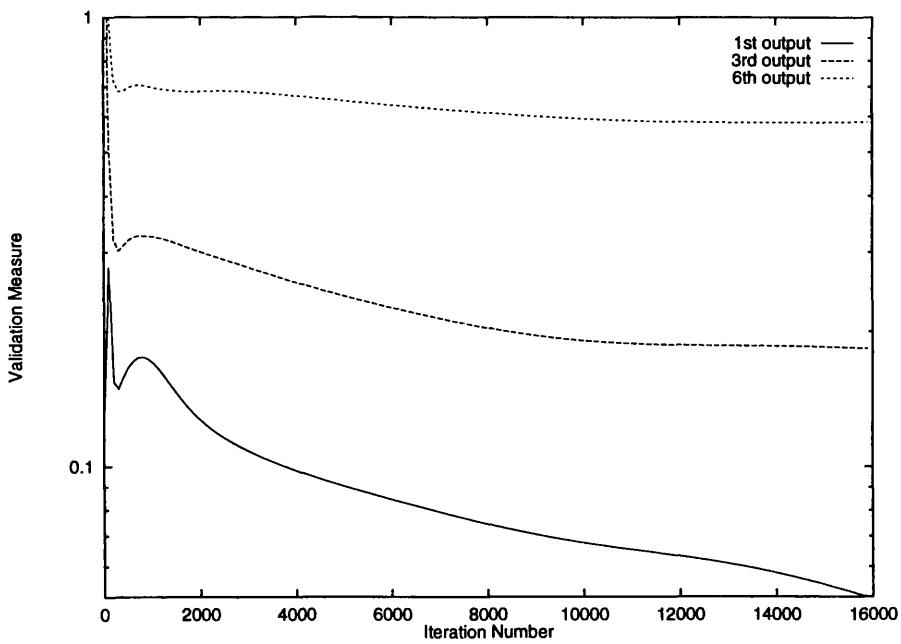
Table 5.7: Neural network prediction accuracy for the smoothed monthly *aa* index. The individual networks and the RMS accuracy are shown for each month's prediction.

other quoted results are not the best obtained, simply that other networks which looked promising on other prediction timescales were not necessarily trained long enough to reach their minimal point in training error.

As mentioned in the introductory section to this chapter, since the *aa* data is more limited than the sunspot data, these networks were trained using internal validation as a monitoring method for training. This was studied in some detail in Section 4.3.2 and found to be satisfactory although not necessarily ideal. For this reason, we study it again here for a couple of example cases. The fact that each output unit of a network generalises best after a different number of training iterations has consequences for the implementation of internal validation (and also the check-set method). This was commented upon in Section 4.3.2 but left until now for investigation. For simplicity, the idea is shown first in Figure 5.11 using a comparatively small network of design 18-6-6. In Figure 5.11(a), the internal validation measure is plotted against the training error for the duration of the network training. The training error behaves as expected in that it shows the network still learning the training patterns. From the behaviour of the *arv* over all the outputs, it would not be obvious that the training should be stopped as the levelling off of the *arv* is very shallow. Table 5.7 indicates however that the best prediction for the 6th output comes after 3900 iterations. For the 3rd output, the best stopping point is found to be after 8900 iterations while on this network, the 1st output had not reached its best prediction, even after 16000 iterations. A glance at Figure 5.11(b) confirms these results fairly well, with the *arv* calculated for the 1st output still decreasing, the 3rd output curve levels off at approximately 10000 while the 6th output also



(a)



(b)

Figure 5.11: Investigation of the method of internal validation for determining the best point to stop training a network. The details of these graphs are explained in the accompanying text.

indicates an earlier stopping point. Thus although the values do not correspond exactly, they do indicate that different outputs will generalise better at different stages and so confirm the idea that they should be monitored separately.

In experimenting with this idea on a larger network with 18 outputs, it was not found particularly useful as the *arv* measure became 'unstable' from the 12th output unit. This means that there were several minima in the plot of *arv* against iteration number and it would not be possible to identify the best stopping point from the curve. Thus in practice an individual validation calculation does not seem helpful for longer ahead predictions. Since the check-set approach calculates the generalisation capabilities over a larger range of data, it is potentially more stable and thus it might be useful to carry out calculations over the check-set for individual output units.

5.4.5 Comparison with McNish/Lincoln

The reasoning behind the use of the McNish and Lincoln method as a comparison measure for geomagnetic prediction was covered in a previous section (5.4.2). As before the accuracy of this method was computed over the same prediction range for the same timescales as the neural network predictions. Figure 5.12 shows the comparison between the two methods. It is seen that this comparison follows a similar pattern to that obtained with the monthly sunspot comparison in Figure 4.3(a) in that the difference in accuracy increases as the prediction time increases. In addition, the neural network accuracy is better than the ML method even at 1 month ahead prediction. It is possible also to quote a comparison between the prediction accuracy achieved here and a result from Kerridge et al. (1989). For the same *aa* index, producing predictions six months ahead based only over cycles 20 and 21, the authors quote an RMS value of 2.7. The 18-6-6 neural network accuracy of 1.9 for the same range compares very favourably with this independent result.

5.4.6 Delay in *aa* Predictions

As was witnessed for the unsmoothed sunspot data at the start of the chapter, it is essential to gauge the usefulness of the results which the networks are producing. All may be well in the learning phase, but what the network is learning has to be confirmed. Hence the predictions of the *aa* index are investigated for any delay or echoing taking place. The same procedure of timeshifting the network predictions positively and negatively against the actual data is carried out and the results plotted in Figure 5.13. These figures show quite clearly a delay in all the prediction timescales which is equivalent to the number of months ahead for which the prediction is sought. Section 5.2.3 showed the importance of distinguishing an echo effect from a delay in the prediction of unsmoothed monthly sunspot numbers. The position looks at first more serious here with a delay evident in all the timeshifting plots in Figure 5.13. Thus full calculations using an echo prediction scheme plus the figures for delayed network predictions are plotted against the

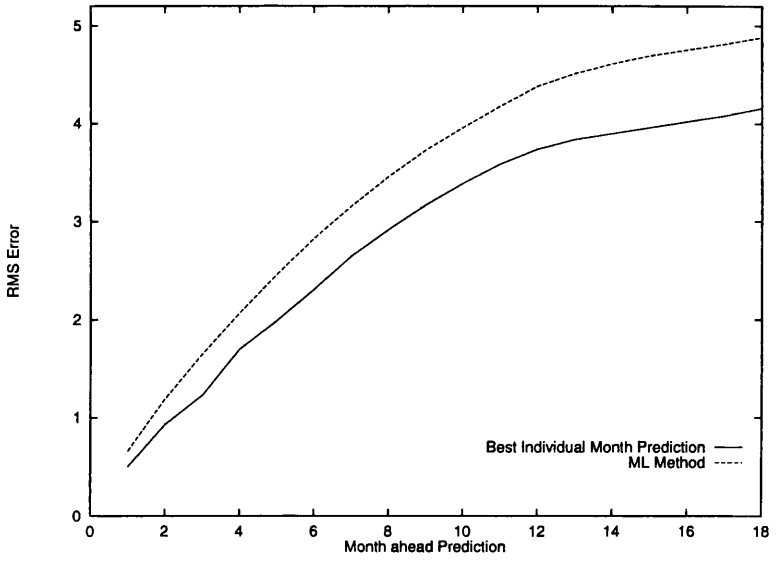
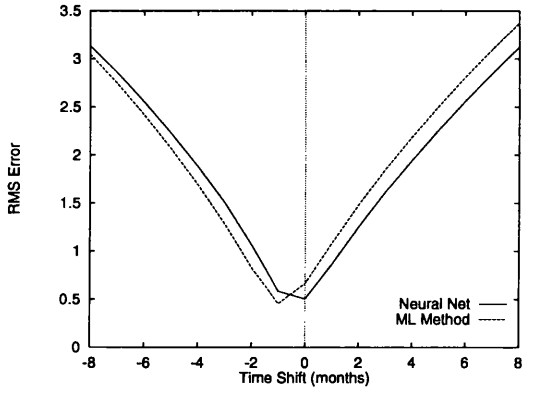


Figure 5.12: Neural networks v McNish and Lincoln method for the geomagnetic *aa* index. The line from the best neural networks is consistently under that of the ML method, showing the superior level of prediction accuracy.

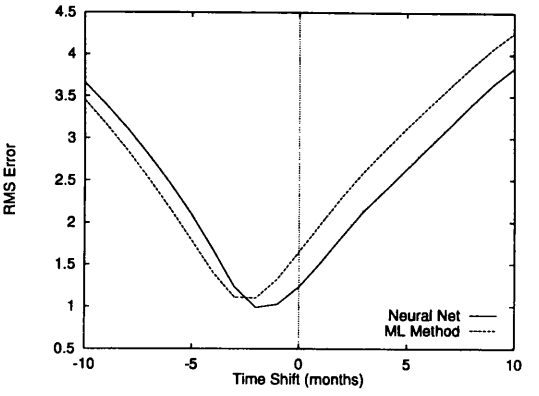
accuracies already achieved for the networks and the ML method. The conclusion again is that the contamination of the results is a delay in the predicted values since a simple echoing method does not provide the more accurate results for each month ahead as suggested by Figure 5.13(f). The RMS Error is greatly reduced at each stage, however, when the network predictions are shifted in time to the minimum of each of the plots. Thus a method of correcting for this delay becomes a future top priority to exploit fully the use of neural networks in predicting this time series data.

5.5 Conclusions

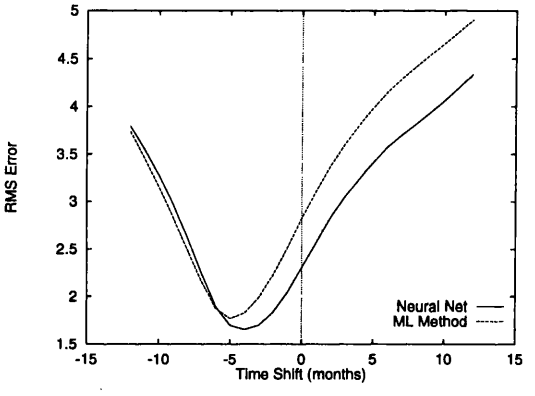
The purpose of this chapter was to extend the range of application of neural network computation to cover other solar activity indices as an alternative to the well-established sunspot number. Combined with this was the ability to study how the networks' are affected when more (or less) rapidly varying time series data is considered. The treatment of unsmoothed monthly sunspot data provided an example of noisier data than was analysed during the first implementation of the methodology in Chapter 4. The motivation for this was twofold in that not only was the networks's ability to learn and predict under study, but also the prediction of this data provides more immediate answers as the value in month m does not rely on the values $m, m+1, \ldots, m+6$ to provide the information to smooth the data. Unsurprisingly, in view of the more complicated nature of this data, the prediction accuracy achieved was well below the results from the smoothed



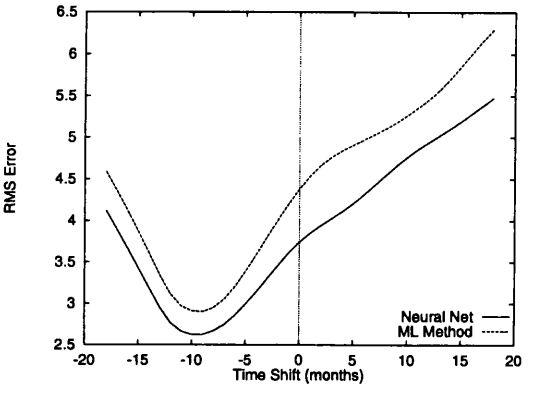
(a) 1 month ahead



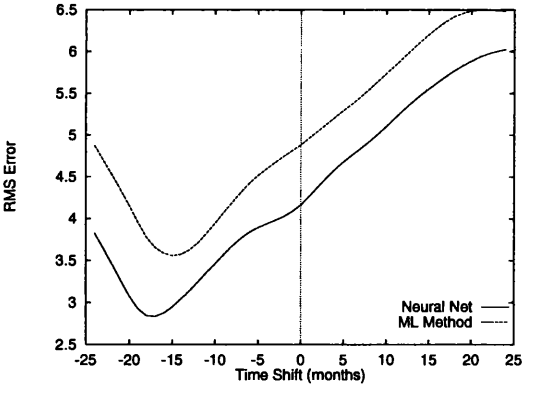
(b) 3 months ahead



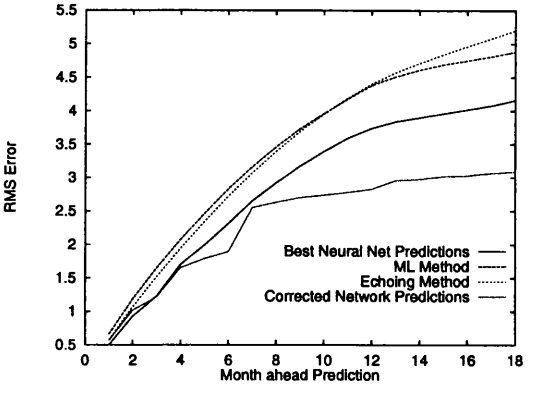
(c) 6 months ahead



(d) 12 months ahead



(e) 18 months ahead



(f) Summary of prediction accuracies

Figure 5.13: Figures (a)-(e) show the time-shifted plots for the network predictions of the geomagnetic *aa* index for the 1, 3, 6, 12 and 18 months ahead. Figure (f) compares the prediction accuracy of the best networks and the ML method with that achievable through an echoing approach or through correcting the networks for the observed delay.

data. The small spread in the RMS Error from 1 to 18 month ahead prediction compared with the spread for the smoothed data indicated a limit on the predictability of this index. This was confirmed when evidence of any delay in the predicted values was sought. The delay which was found for the first few months ahead points directly to the need for a correction method to be introduced to the neural network prediction method.

The importance of the solar 10.7cm flux as an index of solar activity levels was explained in Section 1.4.1. In particular it is used as an input parameter by the European Space Operations Centre to a standard atmosphere model to allow estimations of satellite orbit decay to be made. Predictions of smoothed monthly values and also of yearly values of this index were considered. For the former, two possible approaches were considered using either the sunspot number prediction and using the correlation between the two indices to convert to solar flux predictions or training the networks on an extrapolated solar flux database. The conclusion reached is that the extension of the solar flux values using Equation 1.4 and the subsequent training on this data proved more accurate, although obviously more time consuming. The analysis of time series data where the time interval between points is a year changed the prediction timescale entirely, allowing the first predictions of the behaviour of the next solar cycle (23) to be made. This is a forerunner to subsequent work in the next chapter.

Finally a foray into the prediction of geomagnetic activity variations was attempted. A brief description of the nature of the variations and previous prediction models was described. Although the prediction accuracy of the neural networks trained on the *aa* antipodal index was superior to that obtained using the version of the McNish and Lincoln method in use before, the evidence of a delay effect occurring throughout the prediction range examined here emphasises the need to correct for this. Plotting the 'artificially' corrected predictions (that is, simply shifting them with respect to the actual data) against an echo prediction scheme suggests that echoing is not occurring and so the network is not simply learning to reproduce the last input value. There remains, however, much to be investigated with respect to this delay effect. For example, the timeshift graphs only give an ideal of the average delay over the whole prediction range, but does not address the issue of whether the delay is greater or less than the average value during specific phases such as maxima or minima. One suggested cause is that of sampling intervals, since the input data is sampled monthly while the predictions are made several sample points ahead. The fact that the delay is not the same for the different indices studied, however, does suggest that the complexity of the data is a contributory factor. This is discussed again in Sections 7.2 and 7.3.1.

Chapter 6

Variations of Time Series Prediction

6.1 Introduction

Within this chapter, the aim is to take the basic principles as set out and implemented in the previous chapters and make small adjustments to these methods in an effort to observe the effect on the results which can be obtained. Three variations on the theme are discussed and one different methodology is covered. The first two of the variations have been mentioned briefly before, where the attention turns to *direct multi-step* and *iterated single-step* prediction. In the former the network is trained to learn only one output, but this output maybe 6, 12 or any number of months ahead of the last actual values which make up the input pattern. In iterated single-step prediction the network outputs are fed back into the input layer as subsequent inputs, continuing in this manner so that eventually the network is using only its own predicted values as input data. The stability of this procedure is investigated in some detail. It should be made clear from the start that in what follows, the meaning of *iteration* is as described in the text and is different from previous usage to describe the number of training steps which a network required to achieve its optimal performance. Leading on from this is the second variation. Here a combination of the McNish and Lincoln method and neural networks is considered. The basic premise of ML is to base predictions from the departures from a mean cycle of all past solar cycles. Here we consider using a neural network trained on the *departures* of the sunspot number from a mean cycle (rather than the raw data) to make predictions of the future difference in value of the sunspot number from the mean cycle. Since the mean cycle bounds the network predictions to follow some sort of periodic behaviour relevant to past solar variations, this approach may have more significance compared to the iterative approach first mentioned. Finally as a complete departure from strict time series

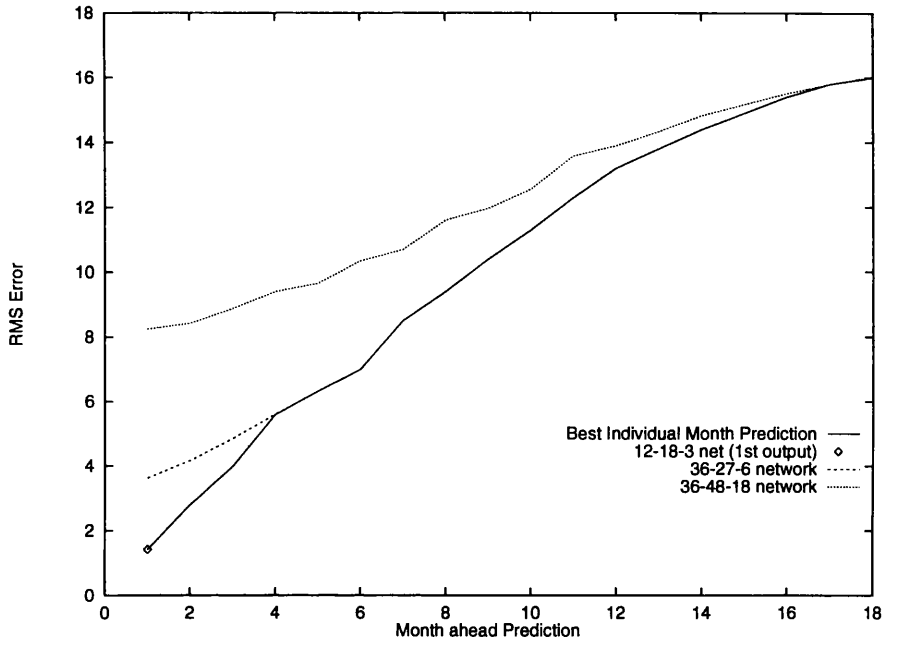


Figure 6.1: The overall best network prediction accuracy for 1-18 months ahead, along with the individual networks for 1, 6 and 18 months. This is reproduced from Figure 4.2.

analysis, the ability of a network to predict simply the time and value of sunspot maximum in a cycle. This idea has already been covered in some detail in Koons and Gorney (1990) and so the work included here is completed as an attempt to reproduce and confirm some of those results. In studying all of these variations we return principally to the international sunspot number as the solar activity index.

6.2 Iterated and Multi-Step Prediction Results

The basic question to be answered in this section can be better understood by consideration of Figure 4.2 and so it is reproduced here for convenience (Figure 6.1) with the addition of the prediction accuracy of the best 1 month ahead network. This figure shows the line of best individual month prediction along with the accuracy of prediction for the best 1, 6 and 18 ahead networks. The best line does not show up as particularly smooth because it includes optimised values for some of the monthly predictions but not for all of them (since only the 1, 6, 12 and 18 values were considered in detail). The plots of the outputs of the different networks show how they are often much worse than the best accuracy available until they approach the particular regime at which they produce their best predictions.

Consider the 12-18-3 network which produced the best prediction accuracy for 1-month ahead. If,

instead of using the 2nd or 3rd output units to get 2 or 3 month ahead predictions, we feed the first output back into the network and recalculate the outputs, this will generate an *iterated prediction* for 2-months ahead. Repeating the procedure will produce an iterated 3-month ahead prediction. The question is: in Figure 6.1 will the accuracy for the iterated 2 or 3 month prediction fall above or below the accuracy already achieved and shown in this graph. Extension of this procedure would lead to similar questions on iterating the first output to compare with 6, 12 or 18 month prediction accuracy. Similarly, the 6 outputs of a 6-ahead net could be iterated once or twice to produce a comparison with the 12 or 18 month prediction accuracy. This description covers the idea of iterated predictions. The alternative proposal of the multi-step approach looks at the possibility that a network with only one output but which corresponds to any point on the x-axis of Figure 6.1 might have a better prediction capability because it does not have to learn all the other points in between.

The terminology which has been adopted in this thesis to describe the different types of approach to prediction is *single-step* and *multi-step* prediction. These terms were already mentioned in Section 3.3 although the consequences of the difference in approach were not investigated at the time. The difference in these definitions is most obvious when discussing neural networks with only one output. The methodology followed so far in this thesis could in some sense be described as a combination of both these methods. This is because architectures with more than one output have been used extensively. Thus in considering a network with n outputs, there is a prediction one time step ahead as well as n steps ahead of the last actual input value (and all time steps between 1 and n). However, since the important difference of multi-step prediction is *jumping* several steps, that is omitting them from any stage of the network, it is assumed that the results to date have been obtained from a single step approach and so examples of true multi-step prediction are tried. The further alternative to these two approaches is the major concern of this section and has previously been termed *iterated single-step* prediction. In this case predictions are made several steps in advance, but using single step jumps, feeding the last network output value into the input layer for the next step.

6.2.1 Monthly Iteration

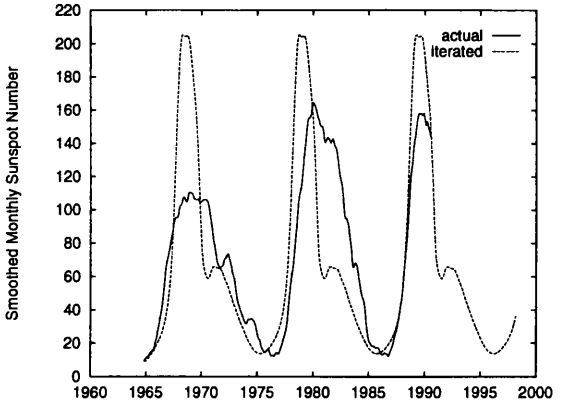
In this section, on iterating monthly values of the sunspot numbers, the aim is to compare the prediction accuracy achieved here with that already established in the earlier chapters. No further network training or learning is involved in carrying out this work since it is only the style of prediction which is being altered. Hence all the previously trained networks with the chosen values for the learning rate are valid and moreover all the training validation results are still in operation as the best point of generalisation is still desired. In essence all that has changed is that when the trained network is presented with unseen data, it is initially only presented with the first pattern.

Feeding these inputs through the network allows the network to calculate its predicted outputs. In the case where $N_{out} = 1$ the only option is to use this output as the last input value for the next pattern and proceed in this fashion to iterate ahead. This can be done for the full prediction range required or alternatively, at some point after m iterations say, the network can be 'corrected' with the actual values before the process of iterating is allowed to continue again. Both cases are considered here. When $N_{out} > 1$, iteration can proceed using only the first output so that although the network has more than one output, effectively only the first value is being used as it is the one which is re-used each time through the network calculations. The key point is that in all of the above, if a prediction is required t months ahead of the last input value, the network has to be iterated t times. It is also possible, however, to feed all the outputs from from a multi-output net back as a subsequent input pattern allowing the network to advance more swiftly to future values of the time series. For example with $N_{out} = m$, for a prediction t months ahead, only t/m iterations are required (or the next integer value above t/m). Again, consideration is given to both approaches.

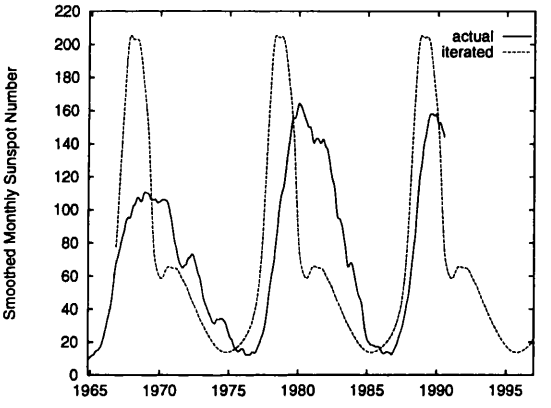
Before we consider the actual results achieved, it is clarified that the aim here is to compare with the accuracies and methods already detailed. The iterated approach will turn up again in a later section (6.4) when the purpose will be to test the ability of this approach to predict the value of an upcoming maximum with a view to allow an early prediction of solar cycle 23.

Results

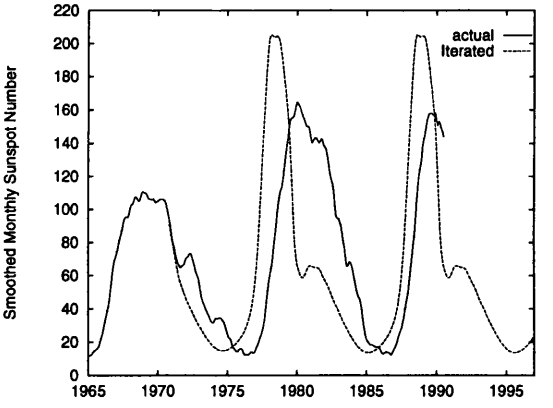
The first examples of the behaviour of an iterating neural network exhibit the behaviour from the first case described above, where the network is only given the initial input pattern and then has to use its own output values for all subsequent predictions. Since it would be natural to assume that the behaviour of such a network would be different depending on which part of a cycle it is given as a starting point, three examples are contained in Figure 6.2. Figure 6.2(a) shows quite clearly what happens when the network is asked to start iterating from the minimum between between cycles 19 and 20. Since most minima have roughly the same value there is little or no information contained in the first input which would particularly distinguish this minima from any other. The network outputs follow a periodic pattern which shows a period of approximately eleven years, the same as it has been forced to learn through training on all the previous cycles. The maximum value of each peak is only slightly larger than the highest value of 201.3 which was met in training at the maximum of cycle 19. This suggests that the main reason the network peaks is because the outputs reach the highest value which the network is used to dealing with (which is only a bit less than the maximum value of 226 which this network can predict). Once the values start decreasing the network is back on familiar ground and can start using its past experience to produce output values. Many of the previous cycles show a feature on the descent phase and this is produced



(a) Iterating from start of cycle



(b) Iterating from rise phase



(c) Iterating from the peak

Figure 6.2: Plots of the best 1 month ahead neural network iterating ahead feeding the output value back into the input layer to generate the next prediction. The three graphs represent respectively, iterating from the start, the rise phase and the descent phase of a cycle to check for any phase dependence. It is clear that no such dependence exists.

Month Ahead	Iterated Output	Multiple Output
1	1.42	1.42
2	3.65	2.79
3	6.21	3.99
4	9.09	5.59
5	12.30	6.31
6	15.67	6.98
7	19.17	8.51
8	22.65	9.42
9	26.11	10.42
10	29.52	11.31
11	32.85	12.28
12	35.93	13.22
13	38.43	13.75
14	40.31	14.37
15	41.81	14.92
16	43.12	15.42
17	44.22	15.77
18	45.15	16.04

Table 6.1: Results table summarising the RMS errors from single-step iteration and the multiple-output predictions of Chapter 4. The latter is clearly a more accurate method of prediction.

consistently about seven years into each cycle.

The question which immediately follows these findings is the obvious one of checking for a dependence on starting phase within a cycle. The starting point of iteration was moved forward until half way through the ascent phase of cycle 20 and the results is shown in Figure 6.2(b). Interestingly the output settles very quickly into the same periodic pattern as obtained before. Finally from the descent phase (Figure 6.2(c)), the outputs follow a smooth curve to a minimum value before the rise phase begins the standard iteration pattern.

Thus when no subsequent updating with actual data takes place, the network predictions relax into a pattern and follow it. We now consider the case where the first output of the network is fed back into the input layer as before, but in this case after 18 iterations, an input pattern comprising actual observed data is fed in, the outputs calculated and then the method of iterating for 18 steps is continued. This is done for the whole of the prediction range of cycles 20,21 and 22 as before and generates an 'iterated' prediction for each of 1 to 18 months ahead. These values can then be compared with the equivalent values obtained in Chapter 4 using the multiple output networks. The RMS Errors for this method of prediction are shown in Table 6.1 along with the comparable values from before. Since the first month predictions are calculated in the same manner, the RMS Error must be the same as the predicted values are the same, allowing a quick check on the correctness of the calculations. Thereafter, the iterated prediction accuracy becomes rapidly inferior to that obtained using the multiple output method previously established. Figure 6.3

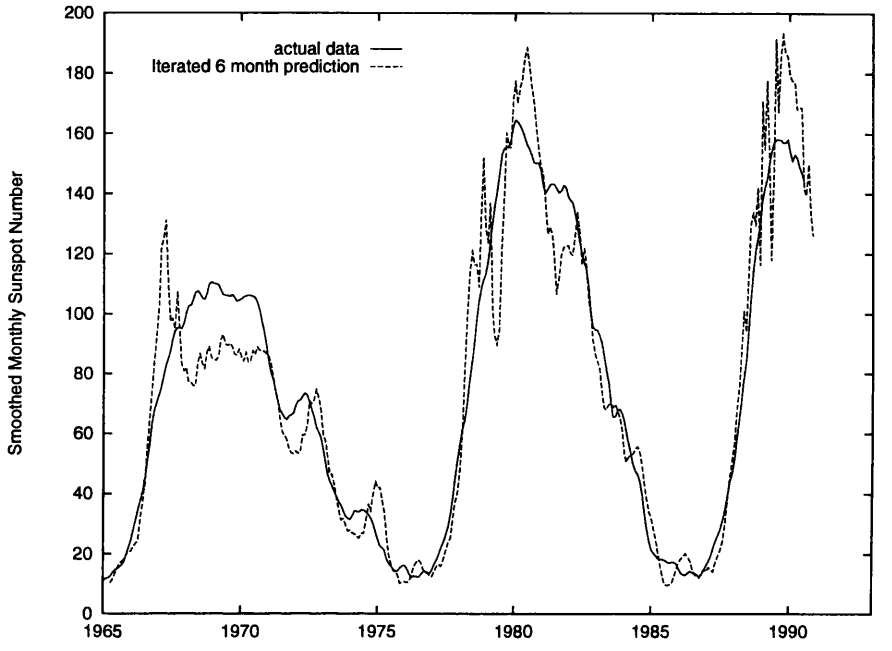


Figure 6.3: The iterated 6 month predictions from the 36-27-6 neural network plotted against the actual sunspot data. It is immediately obvious that this method is not nearly as accurate as the multiple output predictions obtained in Chapter 4.

provides a visual explanation of why this occurs. During the ascent stage of cycle 20, when the network starts to iterate itself, it rapidly sets off to reach the high peak which was established in the case of repeated iteration. After 6 months the network is informed of the actual observed values of the time series and the outputs readjust to the actual data. This free-running state followed by being 'pulled back into line' leads to the sharp changes in direction which are particularly noticeable around the peaks

The prediction accuracies and the plots have only being shown for the case where the prediction range started at the beginning of a cycle. Potentially, it is possible for the success of this approach to vary with the phase of the cycle where it starts. The conclusion from Figures 6.2 that phase is not important suggests however that any improvement which changing the starting phase has on the values in Table 6.1 would only be a local effect and that over a prediction range of more than one complete cycle, the form of results of Table 6.2 would be repeated.

Finally, consideration is made of the case where all the outputs of a multiple output network are fed back into the input layer and the next set of outputs are calculated. This reduces the number of iterations required to reach a prediction n months ahead, thus reducing the error inherent in every pass made through the network. There is, however, more of an error in the input pattern since this now contains all the predicted values from the initial network calculations instead of just

Network	Iterated	Prediction Time	RMS Error
36-27-6	Once	12 months	12.39
	Twice	18 months	27.28
12-18-3	Once	6 months	7.64
	Thrice	12 months	19.88

Table 6.2: Summary of results for iterating all the previous outputs. Iterating once appears worthwhile to compare results, beyond that accuracy decreases rapidly.

one value which deviates from the observed value. Two examples were tried to gain an insight into the effectiveness of this approach. Firstly the 36-27-6 network was iterated once to give a 12-ahead prediction and then again to give an 18 month prediction. It was found that the 12-ahead prediction was more accurate than the previously attained accuracy using the 12th output of a network. Iterating again resulted in a much poorer estimate for the 18 month ahead prediction although this result was still substantially better than the 1 output net iterated 18 times. Secondly, the three outputs from the 12-18-3 net were iterated to give a 6, 9 and 12 month prediction. The six month was worse than 36-27-6 net although not considerably. Looking at a triple iteration to achieve the 12 month results, these were now much worse than the original 12 month multiple output accuracy of column 3 in Table 6.1 although again superior to the column 2 figures. These results are summarised in Table 6.2. These results suggest that it is worthwhile iterating the whole outputs once and comparing the error with that from a multiple output prediction. Beyond that the errors involved in iterating compound and the results rapidly become considerably worse.

6.2.2 Multi-step Prediction

The results in this section were obtained from a subset of all the network architectures used in the previous chapters. Since the predicted value is several months ahead of the last input from the time series, only one output unit is required. This output corresponds to the point $x(t+n)$ where the input pattern takes the same form as expression 3.2 and n is the point number ahead in the time series. The number of input and hidden units were varied in the same fashion as described previously to establish the architecture which proved most suitable for this type of approach. Since most of the optimisation processes carried out for smoothed monthly sunspot number analysis concentrated on 1, 6, 12 and 18 month ahead values, these are the only values which are taken for n at this point. Obviously the case $n = 1$ has already been dealt with and so only the three other cases need to be studied.

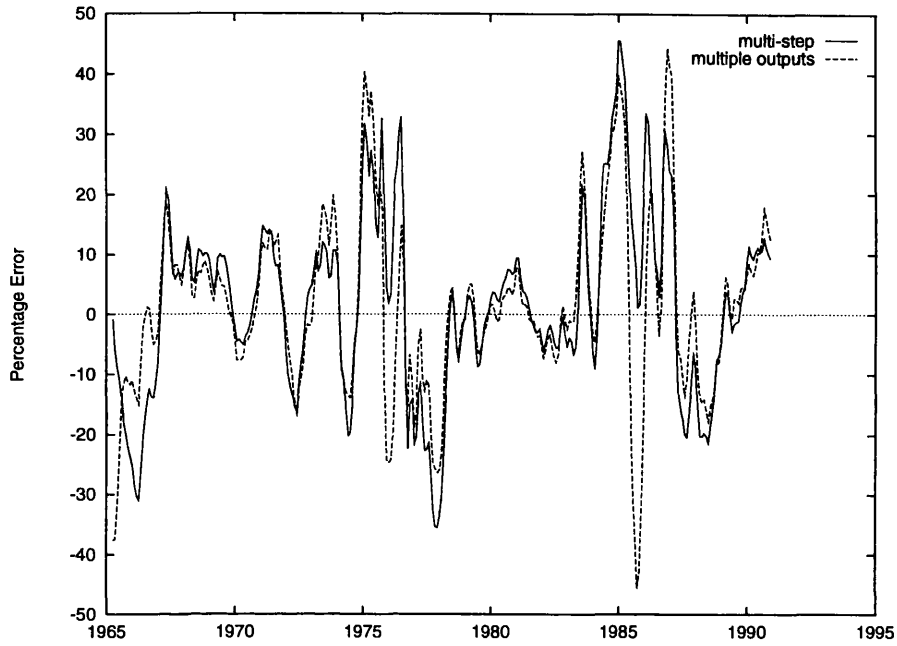
This approach is quite significantly different from the previous descriptions in some respects while not so in other ways. Previously, the idea of time does not appear explicitly in the form of expressions 3.2 and 3.3, instead the fact that the time interval between each point in the input and output pattern is the same (δ, say) allows for recognition of the network predictions for each

Month ahead	Network	Iteration	Multi-step RMS	Multiple Output RMS
6	36-21-1	6000	7.67	6.98
12	36-24-1	2700	13.31	13.2
18	36-24-1	1900	16.15	16.0

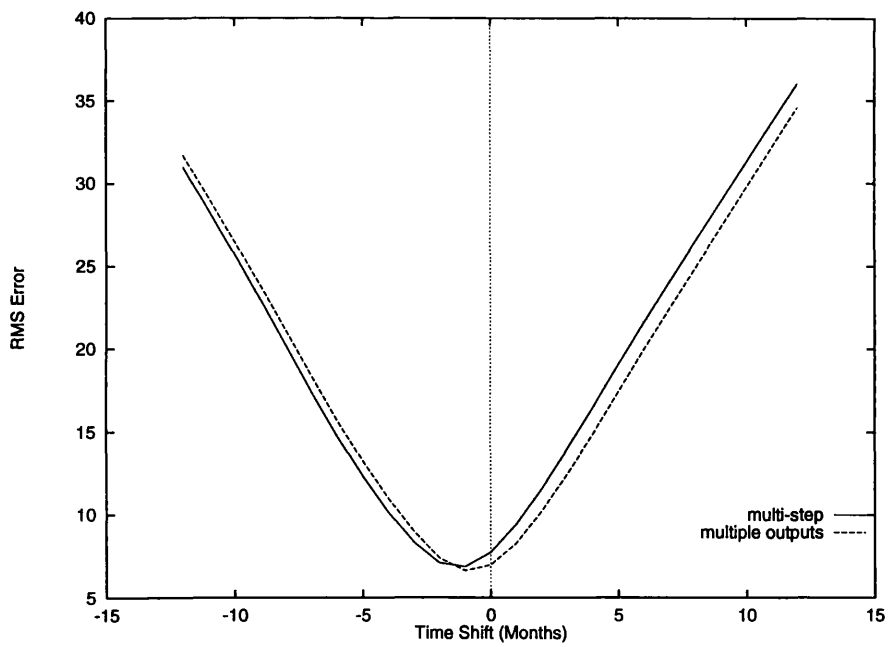
Table 6.3: Comparison between the best multi-step predictions and the errors calculated from the multiple output networks.

output as $1\times\delta$, $2\times\delta$, etc steps into the future. In the case of multi-step prediction, we are effectively assuming that, if δ is the assumed time interval between the input values, then we have $\delta_1 \neq \delta$ as the time interval for the output values, although usually there is only one output. There is little reason to assume, however, that this difference, although formally quite significant, should have any effect on a neural network's ability to learn the output patterns from the inputs supplied since the formalism of expressions 3.2 and 3.3 is irrelevant to network learning. Instead the network produces its own response which simulates that which exists in the data. Thus the exact value of δ_1 and its relation to δ is non-existent to the net, instead it is trained to minimise the errors between the given input-output patterns and its own calculations. The question which does remain to be answered through this analysis of multi-step prediction is which style of presentation of the data does the network find easier to simulate.

For training purposes and determination of the most suitable network architectures, fewer networks needed to be tested since N_{out} is fixed as equal to one. The number of inputs was again tested as 12, 18, 24 and 36 units, with the variation of hidden units from 3 to 36. The learning rate ϵ, momentum α and the number and range of training patterns were all kept identical to the setup which gave the optimal results for predicting the smoothed monthly sunspot number in Section 4.2. This gives as close a comparison as possible between the two styles of data presentation. When the network performances are evaluated using the prediction set the results are found to be slightly less accurate then those obtained using the multiple output networks (see Table 6.3). The differences are not in any case particularly significant and so it is not possible to draw definitive conclusions over which style of data presentation and learning is preferable. Two further tests are carried out to check whether there are any more significant hidden differences between the two prediction approaches. In Figure 6.4(a) the percentage errors of the two styles are plotted for the prediction range, for the 6 month ahead predictions. Aside from two noticeable peaks around cycle 21 and cycle 22 minima, the errors appear similar to each other in this test. In checking for a delay it is found (Figure 6.4(b)) that the multi-step prediction shows a slightly greater delay factor although again it is not particularly significant. Thus we conclude that in this survey the multiple output approach provides the superior performance but the evidence is not strong enough to suggest that this is a general result for neural network time series prediction.



(a) Percentage errors in multi-step prediction (6 months ahead)



(b) Delay in 6 month multi-step prediction

Figure 6.4: The top figure shows the percentage errors in multi-step prediction compared to multiple-output prediction. No significant difference exists between the two plots. The lower figure tests for a delay in the multi-step prediction approach. It is found to be slightly worse than that for multiple-output prediction.

Month ahead											
1	2	3	4	5	6	7	8	9	10	11	12
15.09	20.83	24.67	27.81	28.51	28.64	28.11	28.57	29.04	29.66	30.83	33.39
15.09	19.65	22.72	22.36	21.35	20.06	20.46	20.82	21.14	21.85	25.06	28.67

Table 6.4: Comparison between the predictions of yearly solar flux values obtained firstly from a multiple output (top line) network and then through iterating (bottom line) the first network output.

6.2.3 Yearly Iteration

This subsection considers a similar treatment to the previous ideas except that the networks which were trained on yearly solar flux data are iterated and compared with the direct multi-output approach. The usefulness of yearly iterated values as long term predictors of whole cycle behaviour is left until the example of cycle maximum predictions is considered . Comparing the results obtained through the different approaches of multiple output nets and iteration (see Table 6.4) shows that in this case the iterated approach is more accurate on each of the 1-12 year predictions. Moreover the iterated results show the interesting feature of decreasing from the 3 to 6 year ahead predictions before slowly increasing again. Looking now at the percentage errors which these prediction accuracies translate into, Figure 6.5 shows the percentage errors for the 1, 3 and 6 year ahead predictions. In all cases the error falls within $\pm 25\%$, not particularly good, although in general the majority of the results lie within $\pm 15\%$ which is more acceptable. It is also just noticeable from this figure that the percentage errors in the 1, 3 and 6 year predictions are of the same order as each other throughout the prediction range. There is less tendency for the larger percentage errors to occur at solar minimum, with a more even spread occurring throughout the cycles.

6.3 Combination of Methods (Neural Nets and McNish and Lincoln)

The results of Section 6.2.1 showed that when left to iterate in a free-running mode, the neural network predictions settle into a periodic pattern with a period compatible with the length of the solar cycle but of little relevance to any predictive ability of future cycles. As previously described in Section 1.6.1, the method of McNish and Lincoln is based on the following assumptions:

(i) In a time series exhibiting cyclic tendencies an estimate, to a first approximation, of a future value of the series is the mean of all past values for the same part of the cycle and

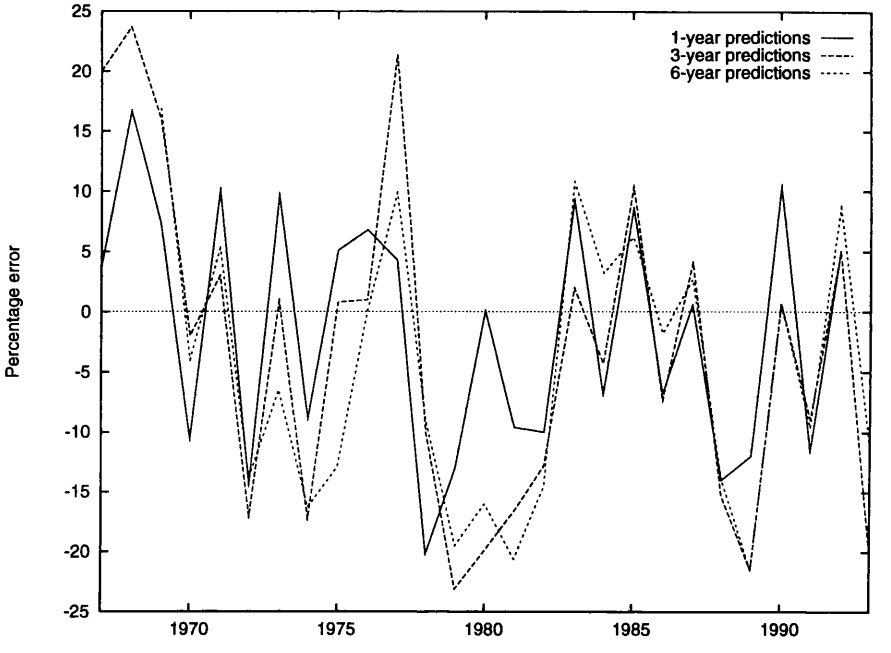
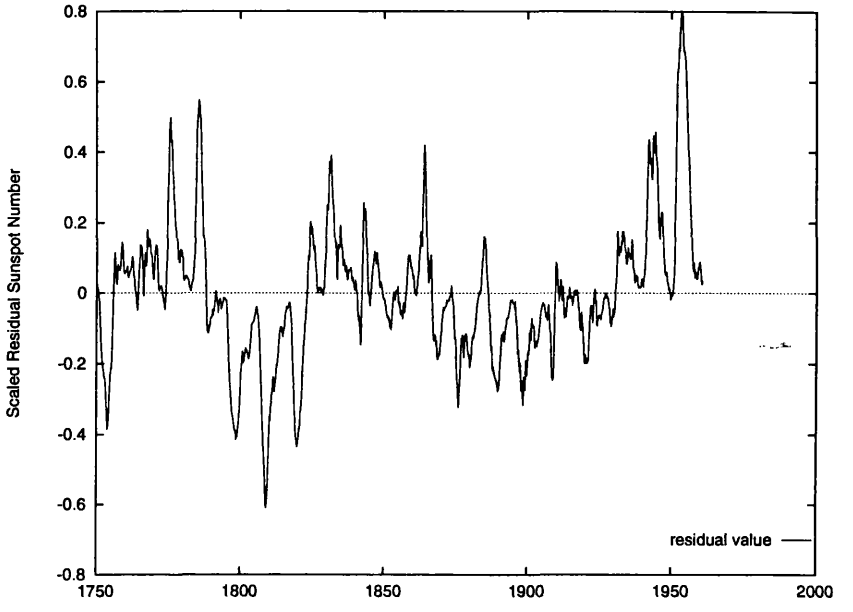


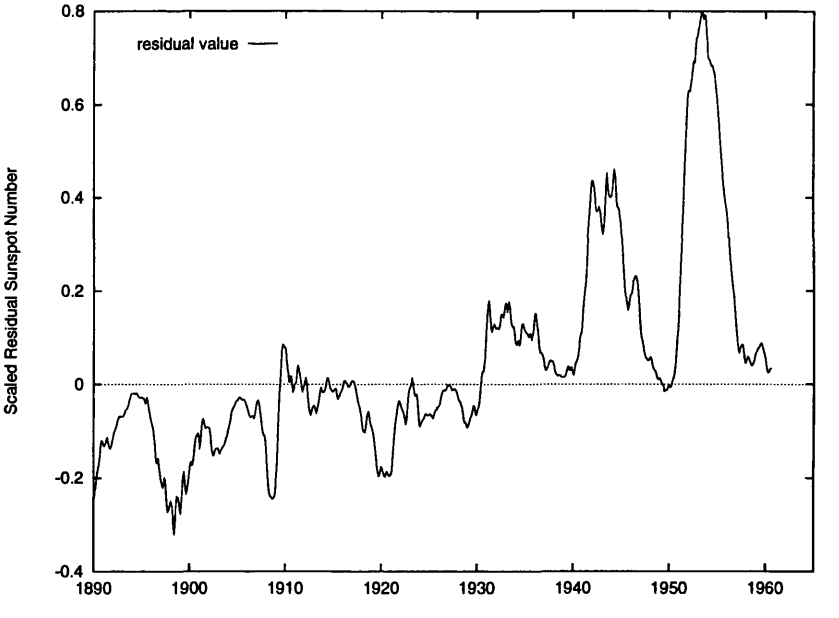
Figure 6.5: Percentage errors for the 1, 3 and 6 year ahead predictions, showing an even spread through out with little phase dependence on any cycle.

(ii) this estimate can be improved by adding to the mean a correction proportional to the departures of earlier values of the same cycle from their respective means, the factors of proportionality being determined by the method of least squares.

In this section the aim is to start off by making the first assumption and so the mean cycle, based on the mean of all past values for each part of the cycle, is constructed first, exactly as carried out during the ML approach. In carrying out the second stage, the method of improving the estimate is changed and the investigation centres on whether a neural network trained on the residuals between the mean cycle and previous cycles can predict the equivalent differences which will occur in the future behaviour. In this way it is hoped that the network is constrained to follow a periodic pattern which makes up a first estimate to the behaviour of the sunspot number. Thereafter it is a question of how well the network can learn to predict the residual components. The most effective way to describe the problem in hand is again graphically. Figure 6.6(a) shows the residual values of the sunspot number from solar cycle 1 through 19 after subtraction of the mean cycle. The most obvious fact about this plot is the fact that it now varies between ± 0.8, thus necessitating a change in the activation funtion of any network trained on this data. The approximate 11-year periodic behaviour of the sunspot number as demonstrated in Figure 1.1 has also been lost, although it is still possible to determine approximately the individual cycles. In particular the negative troughs in the early 1800's correspond to the comparitively low maxima of cycles 5 and 6 while latterly



(a) Cycles 1-19



(b) Cycles 13-19

Figure 6.6: The residual values of the sunspot numbers following substraction of the mean cycle generated by the McNish and Lincoln method. The top figure shows the complete set of cycles 1-19, the lower figure concentrates on the data which was used to train the networks. The obvious differences in this dataset is the ability to fluctuate between ± 1. The different cycles in the training set can still be identified although the obvious 11-yr periodicity is lost.

Month ahead	Network	No. of Iterations	RMS Error	χ^2
1	12-18-3	1600	14.3	724
6	36-27-6	6000	23.2	1333
12	36-48-18	200	28.5	3060
18	36-21-18	500	20.9	1587

Table 6.5: Prediction accuracy for the output values for the combination of neural network and McNish and Lincoln learning algorithm. The results are much worse than those previously obtained through the other prediction methods.

the plot becomes strongly positive as some of the cycles with larger peaks are encountered. The training set for this initial investigation was again restricted to the residuals from cycles 13 through 19. This part of the data is enlarged in Figure 6.6(b). This data is now obviously more complex than the original smoothed data and indeed there is no reason to suspect that any specific trends exist in this data, although the individual cycles can still be identified. It may in fact be the case that the residual time series is chaotic and so this approach gives an initial estimate of the usefulness of neural networks when learning this type of data. This is done for the case of multiple output predictions.

Since we are now dealing with inputs (and outputs) which lie between ± 1 the choice of activation function used in the networks has to be altered. A sigmoidal form is still chosen but with the change of range comes the change from the form of Eq 2.49 to Eq 2.50. This provides the freedom to produce output values in the correct range. Other than this, none of the parameters of the networks were altered for this training. Only the best networks obtained for 1, 6, 12 and 18 month predictions were taken and trained on the residual data during this initial study.

Results

The standard prediction method of using the multiple output units to predict more than one month ahead is analysed. The prediction range was again specified over solar cycles 20, 21 and 22. Since only the specific optimal networks for 1, 6, 12 and 18 months ahead were trained it is only possible to compare results on these prediction timescales. Immediately it is obvious that there is no real comparison when the table of results (6.5) is compared with the RMS and χ^2 errors obtained previously. Although these results were obtained from a fairly brief investigation, the difference in accuracy is such as to suggest quite definitely that it would not be worthwhile to carry out a fuller investigation unless some considerable refinements were introduced to the model. As an example of the sort of results which this method produces, Figure 6.7 shows the various stages which go together to generate the predictions. In the first figure are the predicted residual values, the second figure shows the mean cycle and the third figure is the combination of the first two. Also plotted in the last figure is the actual data for comparison. Although many of the features are correct and

occur at similar times, the height of the curve is generally very much over- or underestimated.

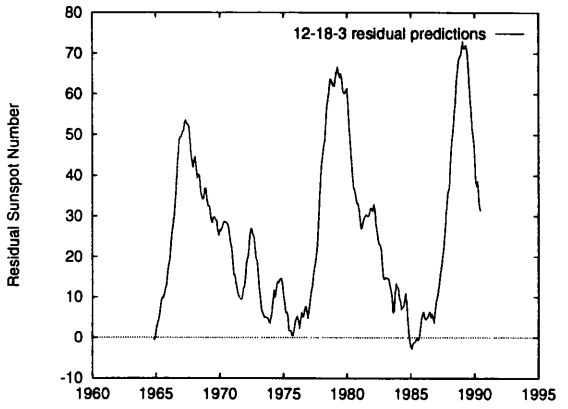
6.4 Prediction of Sunspot Maximum

Koons and Gorney (1990) studied the possiblity of using neural network methods to predict the time and magnitude of the upcoming solar maximum, based on the early onset behaviour of the cycle. This sort of study only aimed at producing two numbers to represent time and maximum, with no interest in the rest of the cycle behaviour. However, accurate prediction of sunspot maximum is an essential part of building up an understanding of solar activity and so the problem is re-addressed here. This is done by first describing the approach and results achieved by Koons and Gorney. Then we attempt to recreate these results by adhering closely to the method described. Finally a couple of minor alterations to the approach will be introduced simply to view any variations in the predictions. These results will be compared with the predictions obtained through the iterative approach.

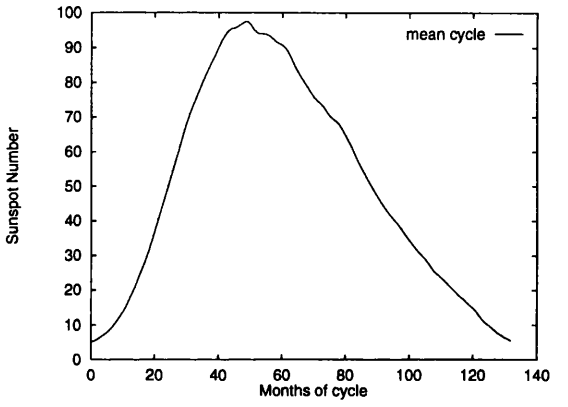
6.4.1 The Method of Koons and Gorney

The neural network used for these results was a commercial package, designed to run on a compatible PC. As will become clear, the size of the training set is small (compared to the time series analysis) and so regardless of the machine or package, time considerations are not particularly important. The training times found by Koons and Gorney are quoted as 15 to 30 minutes. The network used was a standard feed-forward architecture with one hidden layer, with 33, 17 and 2 units in each layer. The 33 units in the input layer were fed input values of 3-month smoothed sunspot number, with the exception of the first input which was given the value of the minimum of the 13-month smoothed sunspot number at the start of the cycle. The 2 output units represent, respectively, the maximum 13-month smoothed sunspot number and the number of months from the preceding sunspot minimum to the maximum value for the cycle. The data upon which training was based was limited to all solar cycles from number 7 onwards, with the previous data presumed too inaccurate for the purpose. Since the principle aim of this work was to predict the cycle 22 maximum, the training set was thus restricted to cycles 7-21 inclusive. Due to the lack of available data, it is not obvious how a check on network convergence can be introduced into the method. Koons and Gorney circumvented this problem by carrying out 15 tests, in each one a different cycle from the set of 7-21 is dropped and used as a prediction test for a network trained on the remainder of this set.

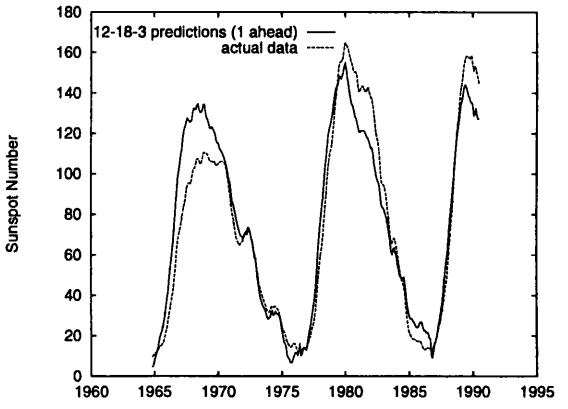
Before attempting to reproduce this method, we summarise the results as described in Koons and Gorney (1990). The training convergence criterion which they placed on their net was 10% of the range. They defined the range, for the first output, as the upper and lower limit of the 13-month



(a) Residual Predicted Values



(b) Mean Cycle



(c) Combined final predictions

Figure 6.7: The various stages which culminate in the prediction of the sunspot number using the combined approach of neural networks working from the basis of the McNish and Lincoln method. Top is the residual predictions, centre is the mean cycle and the bottom figure is the combination of the first two.

smoothed maximum sunspot number for the cycles in the training set (40-210, were the appropriate bounds) and for the second output, the range was 30-80 for the number of months after minimum that the maximum would occur. The range definitions were also used to scale the network outputs between 0 and 1, as required by the network activation function. This puts a training criterion of ± 17 on the sunspot number maximum and ± 5 months on the time to maximum prediction. When the 15 tests were carried out, systematically removing one of the cycles from the training data, the standard deviation of the predictions of the maximum from the actual maximum for cycles 7 through 21 was 26.4. This value was used to place an estimate on the accuracy of prediction of cycle 22 therafter. Using all the available training data, the maximum predicted was 194 ± 26 to occur 42 months (March 1990) after the minimum. Making a prediction of cycle 22 from each of the 15 test cases provided a spread to reflect the uncertainty in the predictions.

6.4.2 Reproduction of Results

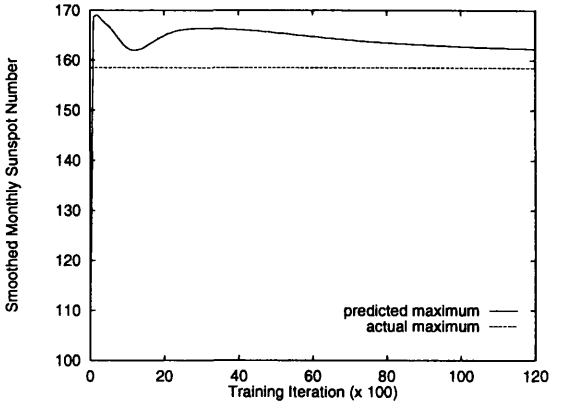
The reproduction of this method which has been carried out here does not adhere strictly to the procedure outlined in Koons and Gorney (1990). In particular, the scaling of the inputs is considered in a slightly different way. The style of using 0.1 to 0.9 as limits for the inputs and scaling only with respect to the maximum value in the time series as used throughout this thesis differs from the the scaling used by Koons and Gorney detailed above. Furthermore it was decided to use all previous cycles for training which provides another factor which could allow the results here to deviate from those quoted above. In addition several facts concerning the exact parameters used in the commercial package which created the original results are not to hand and so the standard network activation function, learning rate and momentum which were used as best initial estimates for the time series analysis are retained. As mentioned above, there is no obvious way of carrying out a check on network convergence during training for this method. There is of course still the inherent danger in neural network learning of overtraining the networks or alternatively for settling for a certain degree of accuracy when more might have been achievable. To help in this respect, a comprehensive set of training objectives were carried out. Firstly a network was trained for 12000 iterations (estimated to be more than sufficient) on all the previous solar cycles and then presented at each hundred iterations with the onset pattern of cycle 22 and the maximum and time to maximum predicted. Additionally a further 21 networks were trained, each one omitting a different cycle from 1-21 from its training pattern. Predictions were made from these other networks for the omitted cycle and also for cycle 22. This will allow an estimate of the errors involved in the predictions to be made.

Since it was the prediction of cycle 22 which was of primary concern in the work of Koons and Gorney (1990) and many other papers of the late 1980's, we consider first the predictions made for this cycle using the full training range. The other nets will then be used to judge the accuracy and

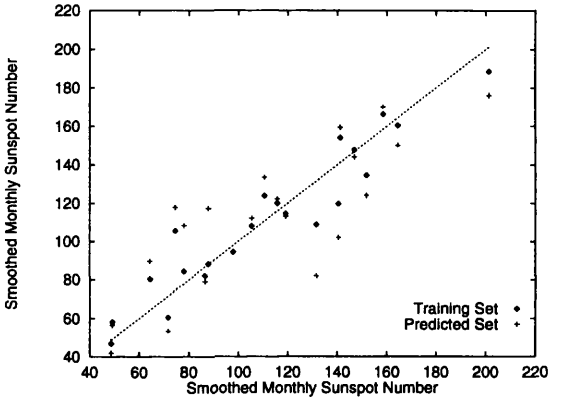
consistency of the result. Thereafter the prediction of the time of maximum will also be considered. Figure 6.8(a) shows the predicted value of the maximum at every sampled iteration throughout the 12000 training steps. The actual maximum of 158.5 is plotted for reference. This shows two potential stopping points during training to obtaine the best prediction. The first is the clear early minimum at 1300 iterations while the steady decreasing trend towards the target at the end of the training iterations suggests a better value still to come. Consideration of the other networks may give advice on which feature is more common to learning this type of problem. In any event, after the initial random phase is overcome, all the predictions are noticeably better than those in Koons and Gorney (1990) (and quoted above) and fall well within the desired error bounds they placed on the training criterion. Figure 6.8(b) plots the predicted values of the training set (cycles 1-21) at the first minimum against the actual sunspot maximum. If the training was perfect, then all the points would fall on the line through the data. This gives an idea of the accuracy of learning at this stage and of the spread of results. Apart from the occasional stray result, the points generally lie within the ± 17 sunspots required by Koons and Gorney and also in general lie within a $\pm 10\%$ bound of the actual maximum value. The other set of points in this figure correspond to the predictions made of the omitted cycle when one cycle is dropped from the training set. In this case the predictions generally fall outside the training set. The third graph in Figure 6.8 looks fairly complicated and unclear initially. It provides a plot of the percentage errors, at each sampled point in training, on the prediction of the maximum of cycle 22 from the other 21 networks trained on one less previous cycle, as well as including Figure 6.8(a) expressed as a percentage error of the maximum. What can be seen clearly from this graph (Fig 6.8(c)) is that there is a definite minimum between 1500-2000 iterations in all of training histories. Furthermore at this point, the percentage error on prediction of the maxima is within $\pm 10\%$. This provides reassurance that this is the best point in the training of these networks to assess their prediction capabilities. This point is used consistently in the following summary of the results.

The prediction results then for the maximum of cycle 22 are summarised in Figure 6.9 where the times and maxima of the cycle from all the 22 trained networks are plotted. The actual values were 158.5, 34 months from onset, which is plotted at the centre of the cross. The extent of the errorbars on this point show $\pm 10\%$ in every direction. It is observed that all the predictions for the maximum fall well within this bound and are closer to $\pm 5\%$ although in general they all tend to overestimate the peak. This cycle peaked very early for a solar cycle and that explains the fact that many of the times to prediction fall close to or just outside the $+10\%$ bound on the time from previous minimum. The predicted values from the network trained on all the previous 21 cycles gave results of 162.1 for the maximum occurring 38 months from onset. This is typical of the general results.

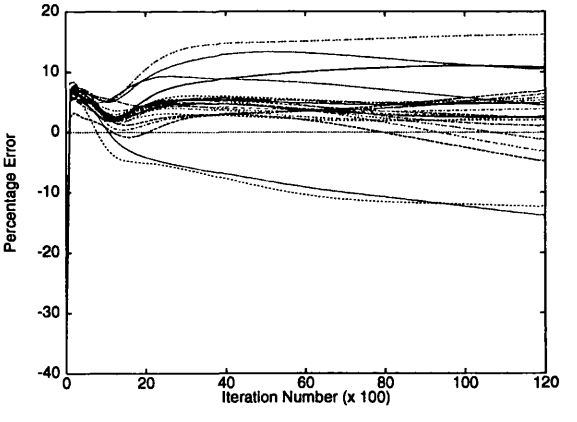
While this section started out to recreate the results of Koons and Gorney (1990), several differences



(a) Cycle 22 prediction based on cycles 1-21



(b) Training and prediction cycles



(c) Training on all cycles

Figure 6.8: These figures represent a graphical summary of the training results while reproducing the method of Koons and Gorney. The individual figures are described in more detail in the accompanying text.

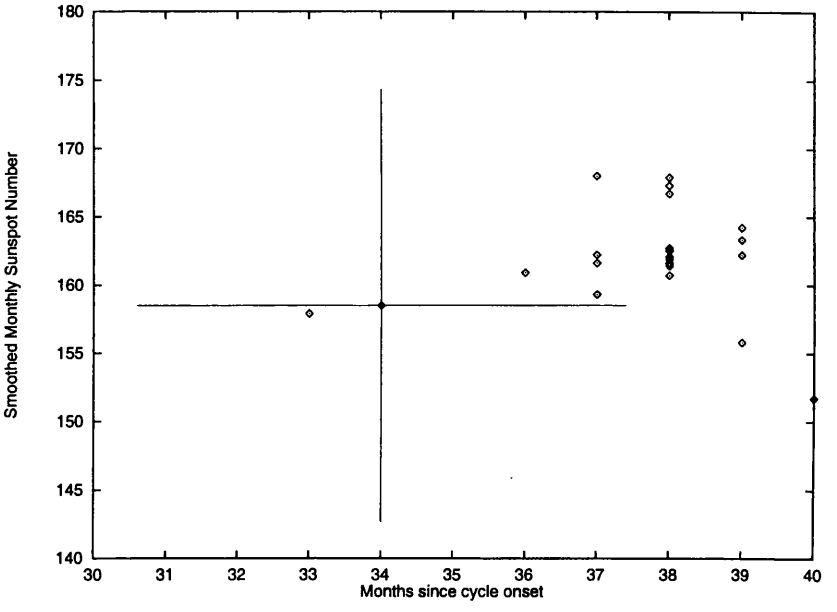


Figure 6.9: The predicted values for time and maximum of solar cycle 22, based initially on the approach of Koons and Gorney.

were introduced in the exact methodology. The results not only confirm the earlier findings but improve on them substantially. Before considering a couple of slight variations on this approach we analyse more closely the ability of these networks to predict the time of sunspot maximum. We take the same stopping point for training that gave the best predictions of the maxima. Figure 6.10 is equivalent to Figure 6.8(b) except that now the x- and y-axes give the time from cycle onset to maximum. As before if the training was perfect all points would lie along the solid line. Again the predicted values for the cycles omitted from the training are included and again these are in general less accurate. Most of the points lie on or inside these bounds although there are several outlying points, including one from the prediction set well below the line. It should be remembered that it is quite possible that the best training point for the time predictions is different from that for the maxima and so this plot could be improved. This is consistent with the results from the earlier chapters that different outputs perform best after different training times. For now it is taken that the maximum value is the more important of the two outputs and so the accuracy of the time predictions is good enough as support for the primary objective. An alternative to this would be to train the networks separately for maximum value and time since onset and put the predictions together at the end.

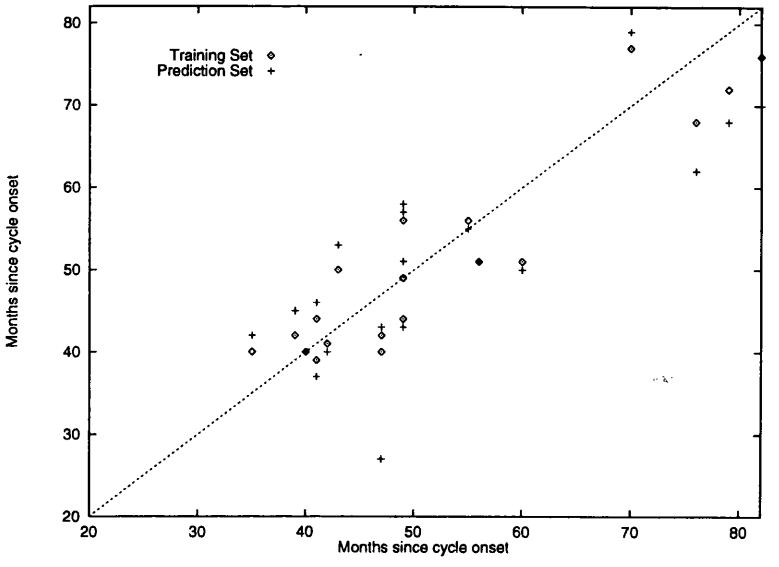
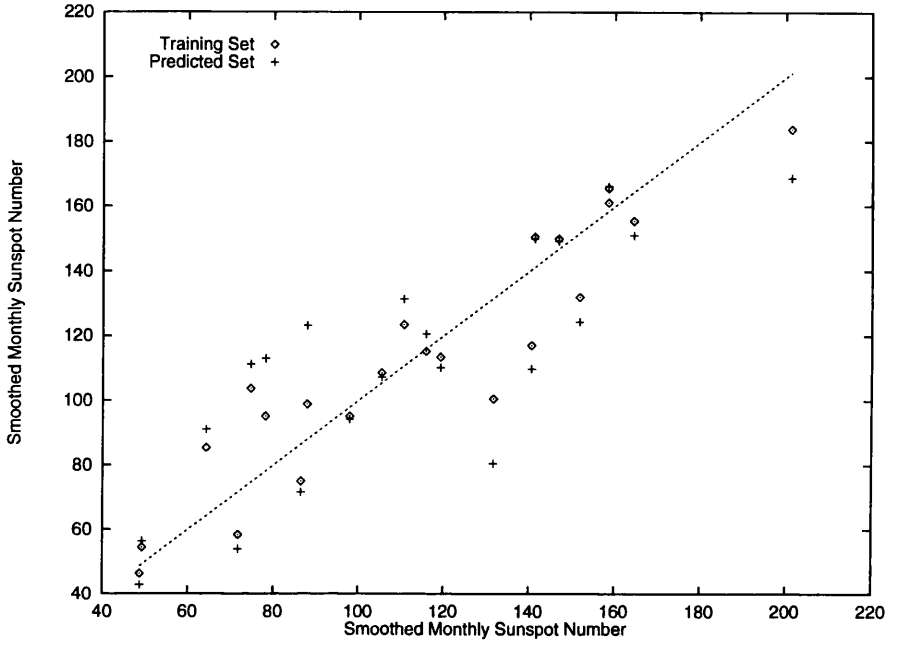


Figure 6.10: Scatter diagram showing the predicted time from onset against the actual time from onset for the occurrence of the maxima of cycles 1-21. The training set represents the results when the whole set is used to predict cycle 22, the prediction set is when each cycle in turn is omitted from the training set.

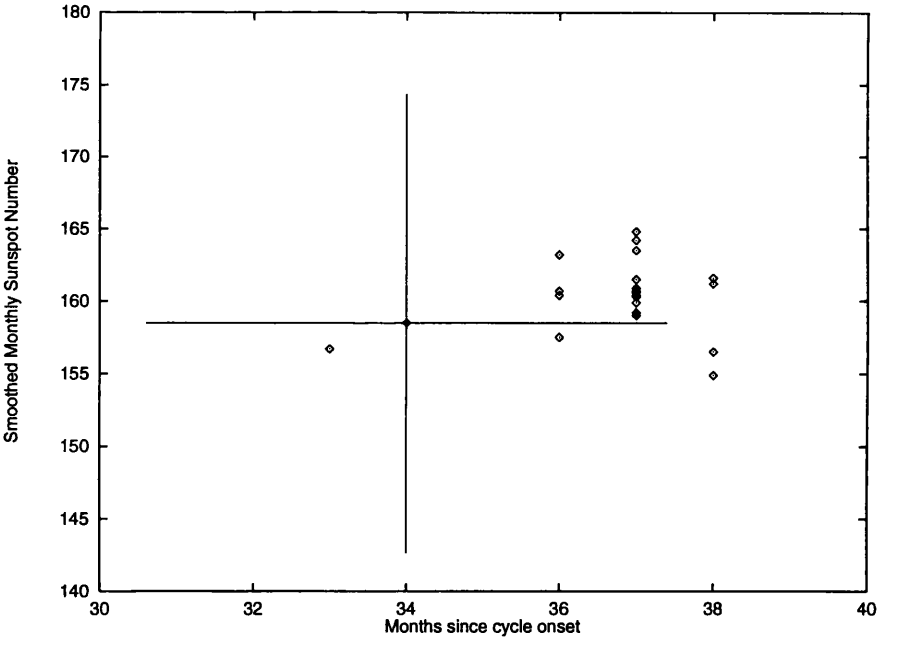
6.4.3 Variations on Section 6.4.1

Two changes to the approach of the previous section are now considered. These changes make allowance for (a) a variation in the number of hidden units and (b) the 13-month smoothed data being used consistently as input data instead of the 3-month smoothed data. These are considered in turn.

(a) It is not thought necessary at this point to carry out such an exhaustive study using different numbers of hidden units as was done for the time series analysis sections. Instead one alternative is considered, lowering N_{hid} to 12, and depending on any significant effect this has on the results may lead to the conclusion that further investigation is required. Carrying out the same analysis as Figures 6.8(a) and 6.8(c) suggests 900 iterations as the most suitable stopping point at which to analyse the prediction accuracy. Plotting the predictions of the training set against the actual values in Figure 6.11(a) gives an indication of a slightly wider spread in accuracy. Looking at all the details of the various prediction files suggested that this was more because the minimum after 900 iterations was shallower than the minimum for the network with 17 hidden units and so it is harder to specify the best point. The prediction accuracy for cycle 22, however, is easily shown to be improved by this network. The equivalent plot to Figure 6.9 is shown in Figure 6.11(b) and it is seen that not only are the predictions of the maximum more closely packed but also that the time of maximum is more accurately predicted.



(a)



(b)

Figure 6.11: These figures are equivalent to Figures 6.8(b) and 6.9 except that in this case only 12 hidden units are used in the networks instead of 17. It is observed that the prediction accuracy is improved with these networks.

Although these results are encouraging with respect to the prediction of solar cycle 22, it is inconclusive whether this network is genuinely superior or just happens to fit this cycle better. In order to settle this point, more understanding would have to be gained through studies of several other networks and closer scrutiny of the training errors and the 20 networks which were trained and tested with one of cycles 1-21 also omitted.

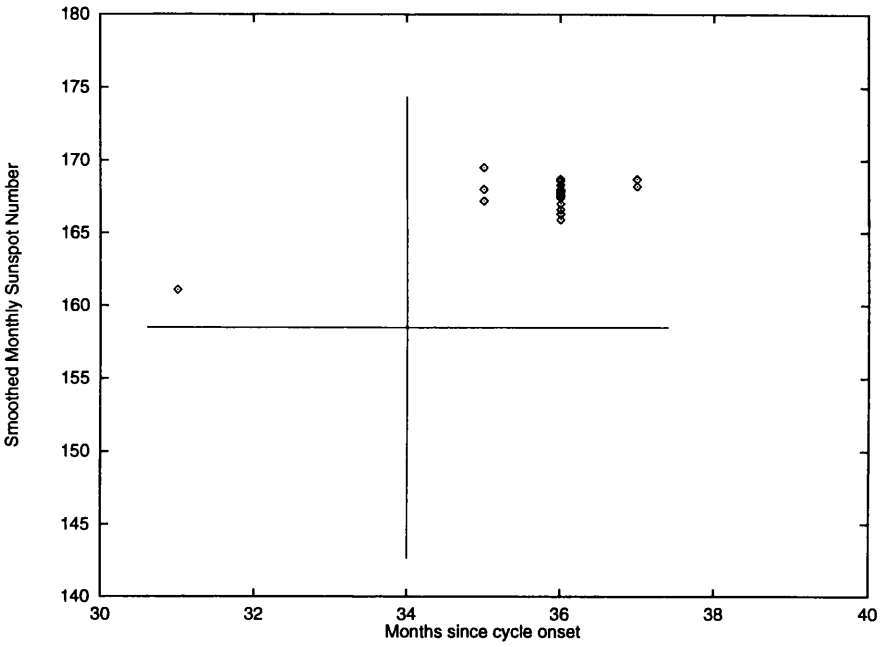
(b) In this variation of the original method, the training data for all the cycles consisted purely of the 13-month smoothed monthly sunspot numbers, without resorting to the 3-monthly data which Koons and Gorney (1990) used. In all other respects the learning phase was identical to that described previously. Again two sets of networks were trained, one with the suggested 17 hidden units and a second set with 12 hidden units for a comparison. In this case only the ultimate results are plotted namely the prediction of time and maximum of cycle 22 (see Figure 6.12). The two sets of networks are kept separate to allow for any trends to be noticed between them. With respect to the results of Section 6.4.1 the prediction of the maximum value is worse for all the networks in this sample. Between the two sets of networks in this section, the nets with $N_{hid} = 12$ do not overestimate the size of the maximum by quite so much as those networks with $N_{hid} = 17$. In all cases the prediction of the time of maximum is within the $\pm 10\%$ bounds quoted before and much more consistently predicted. The most obvious reason for this extra level of consistency with all the predictions is the use of the smoothed data as input.

In terms of fitting the error bounds for both desired quantities, however, it is these networks which satisfy this criteria, since when the 3-month smoothed data was used, the time to maximum was in general far overestimated. Thus it depends on what factor is desired before it is obvious which data or size of network is suggested. This reinforces the idea of training networks separately to predict the time to maximum and the actual value at sunspot maximum.

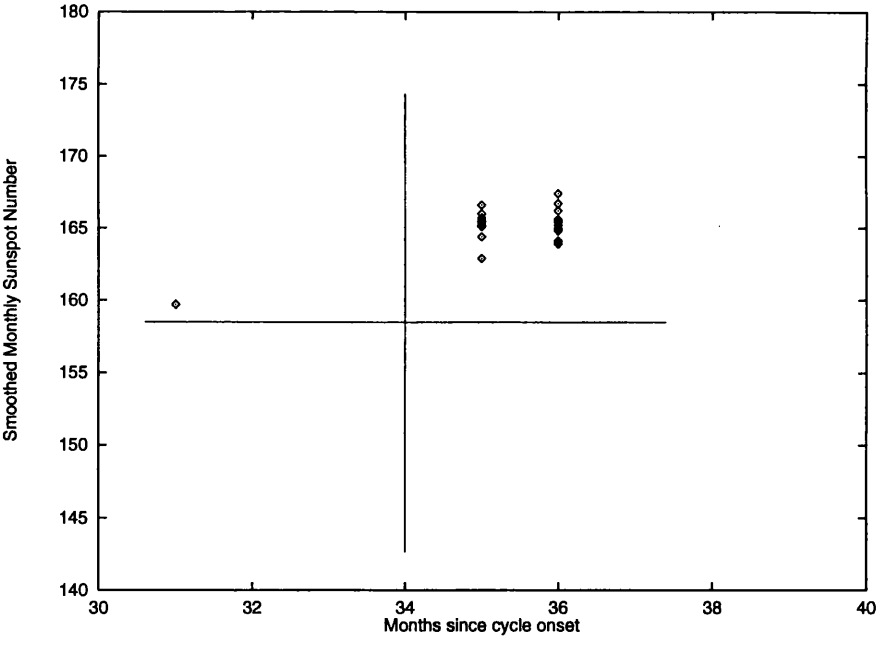
6.4.4 Time Series Predictions of Sunspot Maximum

Finally in this chapter, we return to the time series neural networks and consider only their accuracy in predicting the sunspot maxima encountered in the prediction range. This goes all the way back to Section 3.4.7 where this was suggested as another criteria to determine the best predictions methods for solar activity. The predictions gained from the multiple output networks are tabulated first, followed by an assessment of accuracy when the networks are allowed to iterate towards the maximum.

Table 6.6 plots the maximum values for the prediction cycles 20, 21 and 22 from the multiple output approach and also the ML method. Over the three test cycles, the average percentage error in prediction is 1.2% (1.4% for the ML) for 1 month ahead prediction, 4.4% (6.8%) for 6 months, 12.3% (11.8%) for 12 months and lastly 15.5% compared to 15.2% for the ML for 18 month ahead assessment of the maximum. In particuler the neural network struggles with cycle 20 while the



(a) 17 hidden units



(b) 12 hidden units

Figure 6.12: These are the plots of prediction accuracy for the time and size of cycle 22 maximum using only the 13-month smoothed data. The graphs are split into the networks with 17 and 12 hidden units. Although the size of maximum is predicted slightly worse, the results as a whole are more compactly predicted around the actual values, falling well within the $\pm 10\%$ error bounds shown.

Cycle	Actual Values	Month ahead			
		1	6	12	18
20	110.6 - 11/68	110.3 - 11/68	119.7 - 10/68	134.5 - 10/67	150.0 - 4/68
21	164.5 - 12/79	167.6 - 12/79	167.3 - 12/79	167.6 - 9/79	163.2 - 12/80
22	158.5 - 6/89	160.8 - 6/89	164.7 - 11/89	179.7 - 12/89	174.4 - 3/90

(a) Neural Network

Cycle	Actual Values	Month ahead			
		1	6	12	18
20	110.6 - 11/68	109.9 - 11/68	110.5 - 7/68	117.6 - 7/68	123.7 - 4/68
21	164.5 - 12/79	165.6 - 1/80	171.0 - 12/79	170.9 - 6/80	166.3 - 2/80
22	158.5 - 6/89	162.9 - 8/89	184.6 - 11/89	198.5 - 4/90	210.6 - 3/90

(b) McNish and Lincoln

Table 6.6: The neural network and McNish/Lincoln predictions for time and value of solar maximum 1, 6, 12 and 18 months ahead for cycles 20, 21 and 22 (using the appropriate networks for each timescale).

ML method loses much of its accuracy on cycle 22. The short term prediction of maximum are not of great practical use, so it is the error of $\sim$ 15% which both methods show for predictions 18 months in advance which is the figure of most interest.

As was shown in the previous sections regarding monthly iteration (6.2.1), the results from this method rapidly became considerably worse than the multi-output predictions. Nevertheless, the specific ability to predict the maximum is worth re-examining. In an attempt to gain as much advance warning as possible, the iterations are commenced progressively closer to the start of the cycle, varying from 20 months up until 33 months from onset. The latter figure is bounded by the fact that cycle 22 peaked after only 34 months and so it is of no use to be starting predictions from later in the cycle than the maximum may have occurred. Unfortunately, as can be seen from Table 6.7, the tendency for the iterative approach to lock into the high amplitude maxima which were found in Section 6.2.1 is repeated here. The only conclusion which can be reached from these results is that it is worthless laying any confidence on an iterated prediction of maximum. This rules out the possibility of predicting through this approach the long term behaviour of the sunspot number in, for example, iterating until a maximum of cycle 23 was established.

An obvious question which has to be answered is whether the standard multiple output prediction method of sunspot maximum is suggested over the method described and investigated in Section 6.4.1. The percentage accuracy on prediction of cycles 20 and 21 was respectively 18.8% and 8.8% for the networks which omitted these cycles from their training set. In turn they showed an accuracy of 5.2% and 0.6% on cycle 22. For the network trained on all the previous cycles, the

Month into cycle	Cycle 21	Cycle 22
20	204.1 - 9/79	204.1 - 6/89
21	203.2 - 9/79	123.0 - 2/89
22	204.3 - 7/79	203.1 - 6/89
23	204.3 - 7/79	201.7 - 8/89
24	146.8 - 1/80	202.9 - 7/89
25	129.0 - 2/79	195.9 - 11/89
26	186.9 - 9/80	133.1 - 1/89
27	115.2 - 12/78	190.0 - 8/90
28	114.1 - 12/78	198.8 - 9/89
29	115.3 - 12/78	192.4 - 7/90
30	117.9 - 12/78	194.7 - 5/90
31	128.3 - 4/79	195.5 - 5/90
32	138.9 - 7/79	192.4 - 1/90
33	192.8 - 4/80	190.9 - 3/90

Table 6.7: Iterated predictions for the maxima of cycles 21 and 22. The tendency for these predictions to get locked into reaching the high peak experienced in Section 6.2.1 is obvious again and the lack of stability is empahasised with the occasional spurious result.

accuracy on cycle 22 was 0.7% for the network with 17 hidden units and 0.6% for the net with 12 hidden units. One factor which strongly influences any conclusions is the early occurrence of the maximum of cycle 22. Since in effect the 33 months input to the network gave all the data up to one month prior to the maximum, the prediction accuracy might be thought to be compared with that from a 1-month ahead network. The predictions of cycles 20 and 21 were respectively 16 and 9 months ahead, however, to give their percentage error. It is suggested then that before any solid conclusions can be reached, the method of Koons and Gorney has to be studied more closely to look for any relationship between the prediction accuracy and the time remaining to maximum after the 33 input data points. One way of achieving this would be to assess the prediction accuracies achieved when the number of input units in this method is progressively reduced. In this way prediction of the maximum must be made from less and less information about the onset phase. This work has already been initiated although no results are presented here.

The results above regarding iteration are restricted of course to iterating the monthly **sunspot numbers**. From Section 6.2.3 it was found that the predictions for the **solar 10.7cm flux** on the scale of years were improved through iterating methods. Also in Section 5.3.2 the first effort at predicting the maximum of cycle 23 was made. This objective of long term prediction is completed here with the approach of free-running iterations to produce the upcoming yearly flux values. In order to produce an estimate for the maximum of the next solar cycle (23), the network must iterate for at least 8 years until it would be expected that the next maximum might be reached and be able to be identified positively as the maximum. Although the available data is limited to provide any confidence level on an iterated prediction, Table 6.8 lists what is available from the previous tests on yearly iterated predictions. It considers the estimates for the maxima of cycles 21 and 22 produced

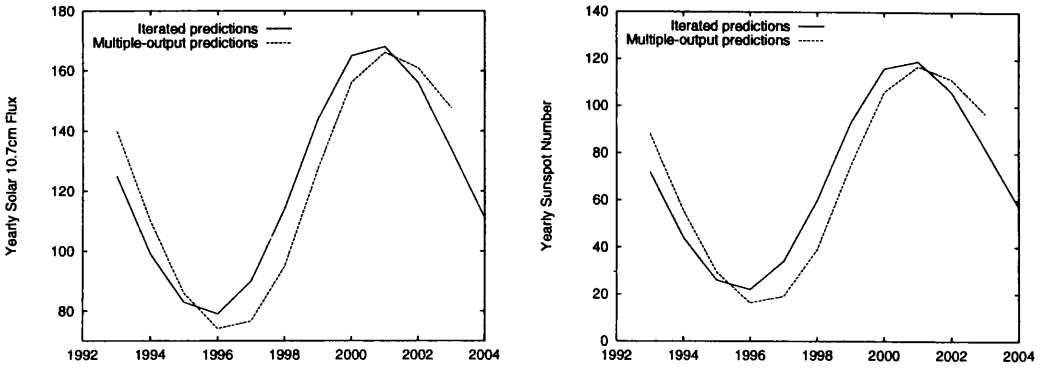
Cycle	Act. Max	8 yrs	9 yrs	10 yrs	11 yrs	12 yrs
21	202.6 - 1981	166.3 - 1980	165.3 - 1980	167.4 - 1980	161.4 - 1981	164.3 - 1979
22	213.6 - 1989	190.0 - 1990	190.5 - 1990	190.2 - 1990	189.5 - 1990	184.6 - 1991

Table 6.8: Iterated predictions for cycles 21 and 22 using the 18-12-12 network. The columns headed by year numbers shows the length of time the network was iterating before it was updated with new data.

Maximum 23	175.9	172.7	167.9	163.8	162.8	166.9	166.5
Date of Max.	2001	2001	2001	2001	2001	2001	2001

Table 6.9: Estimates for the time and value of the maximum of cycle 23 in terms of the solar 10.7cm flux.

after different lengths of iteration and compares with the actual values. As is seen from this table, it is not possible to draw any firm conclusions about the reliability of this type of prediction. Concentrating on the values for the maxima, not the dates, cycle 21 is consistently underestimated by 20% while for cycle 22 the error is very near 10% every time. Starting iterating the 18-12-12 network using the available data up until 1992, generates several estimates for maximum 23 from the different output units. All the estimates are quoted in Table 6.9 and are seen to be fairly consistent within each other with the year of the next maximum predicted as 2001 all the time. Taking an average value from these predictions gives a value of 168 for the next maximum, equivalent to a yearly sunspot maximum of 119 (reconstructed from Equation 5.1). Comparison with the results obtained in Section 5.3.2 using the multiple output method with the 18-12-12 network to predict the next 12 years of solar flux values gives a close degree of correspondence between the two sets of estimates. Figures 6.13(a) and 6.13(b) show the predicted solar flux and sunspot numbers through the next 10 years which cover cycle 23. The solar flux predictions are made using both the multiple output method and iterated values, while the sunspot numbers are calculated from the solar flux values. While the uncertainties in each individual method have been shown to be quite high, they do show close agreement for predictions of cycle 23 which adds some weight to their reliability. In order to produce better yearly estimates of the monthly sunspot number, it would be suggested to retrain the networks on the yearly sunspot data rather than using Equation 5.1 to produce sunspot estimates. This completes this section and so the chapter is concluded with a summary of the results which have been obtained in all the different aspects of this chapter.



(a) Yearly Solar Flux

(b) Reconstructed Yearly Sunspot Number

Figure 6.13: Neural network predictions for the next solar cycle in terms of yearly solar flux and sunspot numbers. The two methods of prediction are in close agreement, adding some degree of reliability to the predicted values.

6.5 Conclusions

This chapter has covered several different ideas and variations on the methodology established in the previous chapters of the thesis. As intimated at the start of the chapter, these ideas fell into four main categories. The first was the consideration of iterated predictions and the second the related idea of direct multi-step predictions. From the results and conclusions drawn in the respective sections, it was found that the errors in prediction magnify considerably with each iteration (that is each sweep through the network). Consequently for all the monthly predictions when the first output was iterated, either in a free-running mode or when they were corrected with the actual data after n steps, the accuracy decreased so rapidly after the first iteration that there was no contest between this approach and the results obtained in Chapter 4 for predicting the smoothed monthly sunspot number. One of the areas where iteration did occasionally work for this index is when all the ouputs of a 3 or 6 output net were iterated once to give a 6 or 12 month prediction respectively. In this case, iterating just once, the accuracies obtained were of the same order or even better.

Moving to the iterations of the yearly solar flux values, this was the other area where these results were better than those of Section 5.3.2. This had consequences for the later section where iterated predictions of the cycle maximum was considered. The result of the investigation into multi-step prediction suggested that the networks did not favour one approach strongly over the other. Direct multi-step predictions dedicate the one output of a network to a specific month ahead value, omitting to learn the values in between. In this study, the results from the best multiple output nets were all better than the multi-step predictions but not considerably so.

The method of combining neural network training on the residual values from the mean cycle of the McNish and Lincoln technique was not studied in full detail. The initial results, however, were significantly worse to the extent that a follow-up on this method would be unwarranteed without significant alterations to the method.

Finally was the rather lengthy section devoted to the prediction of sunspot maximum. This compared the method originally discussed in Koons and Gorney (1990) with a close reproduction carried out here. Furthermore, extensions and variations were also introduced to provide further predictions of the sunspot maximum. The time series analysis approach of Chapter 4 for analysing the smoothed monthly sunspot number was resurrected with a view to check the accuracy of estimation of the maxima and to allow comparison with the approach based on the rate of increase since cycle onset. The subject of sunspot maximum prediction has always been an intensely debated subject and the results of Section 6.4 can only add to the debate rather than settle any of it.

Neural network learning of the onset stage of a cycle in order to predict the maximum was attempted in a follow-up to the original work by Koons and Gorney. For cycle 22 this was found to provide very accurate results in terms of size and time of maximum. This has to be tempered by the fact that the network was provided with all the data right up to the month preceding the point of actual maximum. This is due to the fact that this method is based on using a fixed amount of data from the start of the cycle and 22 peaked very early. In order to compensate for this, it was suggested to reduce the amount of data given to the network and assess the effect on the networks' ability to learn the training cycles and produce predicted values. It is also the case that in many of the training and test examples where another cycle as well as 22 was omitted from the training set that this method was able to produce very accurate estimates of the maximum even when it was many months away.

Other methods of predicting sunspot maximum were also considered, returning to the options of standard multiple output predictions and also iteration methods. The latter proved equally poor for this as for general sunpot number analysis. It was, however, possible to put some estimate of accuracy on the former. Since maximum prediction is only useful if it is provided at least a year in advance, the prediction accuracy of $\sim 12\%$ and $\sim 15\%$ for 12 and 18 months ahead provide some basis to judge this method by. It is worth noting that although the RMS errors over the whole prediction range showed the neural networks much superior to the McNish and Lincoln method on these timescales, in terms of maximum prediction the accuracy was very much closer, with the ML edging out the networks on both prediction timescales. Coupled to this assessment of accuracy is the problem of accurate time of maximum. In very few cases where the maximum was predicted at least a year in advance was the date of maximum very well assessed, even if the size was estimated very closely. Thus while these methods provide areas of further study to provide a closer assessment of the factors influencing accurate prediction of sunspot maximum with sufficient advance warning,

they do not currently appear to provide any significantly greater accuracy of maximum prediction than is already available.

Finally, longer term prediction of the next solar maximum was attempted using networks trained on yearly solar flux data. Since iterated values were found to be of use with this data, these were used along with 12 output networks, to produce a predicted value of $\sim$ 167 for the maximum of cycle 23, to occur in 2001. Using this value to calculate the yearly sunspot number, gives a prediction of $\sim$ 119 for the latter index, indicating a cycle similar in size to solar cycle 20. The more limited amount of yearly data available to provide training and prediction sets makes it difficult to place any firm degree of accuracy on these predictions. Over the three test cycles, the accuracy was $\sim \pm 25\%$ and so this is quoted to give an idea of the uncertainty in the cycle maximum predictions. It is the author's opinion that in any case the best that can be hoped from in this situation is an order of magnitude calculation which would suggest, for example, whether an extremely large amplitude cycle would occur. Even confidence in a method which gave this level of predictability would provide knowledge which would be very useful with respect to the planning of many space science missions.

Chapter 7

Conclusions and Future Work

7.1 Introduction

This chapter splits easily into two distinct parts. First is an overall conclusion to the results obtained and described in the previous chapters. This will draw heavily on the conclusions sections of the individual chapters, while allowing an overall picture of the scope of the work carried out to be given. An important feature of the conclusion sections already written has been to assess the success of the techniques described while keeping close track of any limitations which became obvious. This leads to the second purpose of this final chapter, which is to point the directions in which this work could be extended. These will be in the form of any notable omissions from the work enclosed but will also discuss some more speculative ideas and initiatives.

7.2 Overall Conclusions

The implicit question asked at the beginning of this thesis concerned the applicability of neural network computation techniques to the problem of analysing and predicting the solar activity cycle and the level of success (or failure) which these methods achieved. The purpose of this current section is to provide a summary of the answer which is contained in Chapters 4 through 6. It is the author's opinion that these techniques have proved themselves at least as successful, if not better, than other current prediction models and thus worthy of further investigation as a tool of research within this subject, as well as other potential applications in wider aspects of astronomy.

The above statement represents the simple zero-order answer to the initial question and does not resolve itself into considering the different aims and objectives required of a prediction model and also the different indices and timescales which require to be analysed before the method can be considered acceptable as a prediction tool. Hence the answer is now broken down further into several constituent parts. In the course of this, some additional questions are asked to which

answers are supplied or suggested. These questions relate in general to the implementation of neural networks in theory and in practice.

Following the opening introductory chapter on the nature of solar activity, Chapters 2 and 3 provided an introduction to neural network computation. Chapter 2 covered the theoretical aspects of the subject, including the derivation of the back-propagation of errors learning algorithm, while Chapter 3 was more concerned with the development of the methodolgy for time series analysis using neural networks and the software which was written during this research. The network training and prediction programs were described, including the precise implementation of the learning algorithm, through the text and flow diagrams provided.

The main results and conclusions of this thesis were contained in Chapters 4 through 6. In Chapter 4, all effort was concentrated on the smoothed international monthly sunspot number as the example index for the network performance studies. This is because this dataset is the most complete solar activity index available (see Section 1.3) and also the one used most frequently in prediction models of the maximum level of activity of a solar cycle (see Section 1.6). The success of the neural network calculations as a method of predicting the smoothed monthly sunspot number n months ahead was assessed using two error measures, the root mean square error and Pearson's chi-square error, concentrating particularly on the former. The method of McNish and Lincoln was described and implemented as a comparison measure, representing as it does, one of the most often used prediction methods (including variants of the original technique). It was shown in Sections 4.1.2 and 4.2 that a neural network was able to outperform the ML method over the full range of monthly predictions studied (1 to 18 months ahead inclusive) achieving a significantly lower RMS or chi-square error in every case. These errors translated into $\pm 5\%$ accuracy for 1-month ahead prediction, $\pm 10\%$ for 6-monthly, $\pm 20\%$ for 12-monthly and $\pm 25\%$ for 18-month prediction, excepting occasional larger spikes in *percentage* error during sunpsot minimum phase when a fairly small difference in terms of sunspot number magnifies when expressed as a percentage. Importantly, however, it is at the maximum phases of cycles where it is most important to achieve good prediction accuracy and it is through these points that the percentage errors quoted are easily obtained. Thus the level of ability of neural networks as a predictor for future values of the smoothed monthly sunspot data is established, based on a training phase of the previous history of the time series followed by a test on a completely unseen interval of data.

The results quoted were the best obtained after several processes of optimising the performance of various neural networks. Section 4.5 summarises the various stages of this work, discussing the importance of determining the neural network architecture (that is number of input, hidden and output units) which proves most suitable for each month ahead prediction, as well as the problems which may be encountered if this is not done correctly. It is unnecessary to repeat all of the conclusions from that section. It is sufficient to stress the necessity to gauge the success

of neural network training not on any aspect of the training error but instead on a measure of the generalisation qualities of the network. This is particularly crucial when noisy data is under study rather than noise-free computer generated data and when the network is used for predicting outwith its specific training patterns.

Moving onto Chapter 5, the aim was to implement the knowledge gained through the analysis in the previous chapter while examining alternative solar activity indices or different timescales of variation. The first variation under study was the use of the raw unsmoothed monthly sunspot data. As shown in Figure 5.1 this added a further complexity to the variations in the time series data to be learnt. The most important point to come out of this study regards the emerging existence of a delay effect which degrades the network prediction accuracy. Although the prediction accuracy was still better than the ML method could achieve, there is the suggestion that the network is not achieving all that it could. The delay in the predictions means simply that the minimum error does not occur at the desired month ahead prediction; instead when the predicted data is timeshifted against the actual data a minimum error is obtained at a point several months delayed from the desired point. As was mentioned in the relevant conclusions section to Chapter 5, it remains an important problem to attempt to overcome this delay effect so that the predictive ability of these networks can be improved. It is worth noting that when the input data was simply echoed as outputs this produced a worse prediction accuracy and thus was not the exact cause of the problem. This aspect of predictions will be discussed again in the future work section (7.3.1).

The solar 10.7cm flux has recently become established as a more quantitative measure of the level of solar activity (Section 1.4.1) and so it was important to establish the predictability of this index. Since there exists a close correlation between this measure and the sunspot number, however, it is not surprising that similar results were obtained when studying the smoothed monthly solar flux. Again the neural network results were more accurate than the benchmark test provided by the McNish and Lincoln method. The solar flux values are used by the European Space Operations Centre in a standard atmosphere model to allow for calculations regarding satellite drag and possible re-entry. These results will be used as part of an ongoing investigation contracted to the University of Glasgow by ESOC to provide a report on the applicability of neural networks to predict this solar activity index.

Since it is also important to study the longer term behaviour of the solar cycle, with a prediction timescale longer than eighteen months, yearly data was considered as this dramatically cuts down the size of network which would be required to produce predictions over these longer periods. Consequently the amount of data available for training is greatly reduced also. Extrapolated yearly solar flux values were used for this study and it was found that for the first three years prediction, the ML method was more accurate. The results of iterated prediction in Chapter 6 improved the network results although not sufficiently to be more accurate than the ML for the first

two years prediction. One of the main applications of yearly predictions is in producing long term estimates for the next solar maximum value, but this area of interest (predicting the maximum of a cycle) was left until Chapter 6.

The final aspect of Chapter 5 drifted away from solar indices and paid attention to the observed variations in several of the geomagnetic which have been measured consistently for many decades. In particular, the antipodal *aa* index was analysed. The ML method was again established as a basic comparison technique through the work of Kerridge et al. (1989). In all cases, the neural networks were found to be more accurate, including a comparison with the independent results of Kerridge et al. (1989). The question of the delay effect does however return. When carrying out a timeshifting analysis, the minimum RMS Errors were found not at the desired month ahead prediction but instead were delayed by almost exactly this amount. The accuracies were compared to a straight echo of the last input values but this found to produce less accurate results than obtained initially. This rules out a straight 'echo-effect' taking place but raises further questions about the nature of the delay. This issue is potentially not as simple as may appear and further work will be discussed in Section 7.3.1.

The last part of this overall summary concerns the conclusions which were reached through the analysis outlined in the previous chapter. This involved variations in the presentation of the data as well as specific concentration on forecasting the long term behaviour of the solar cycle through the best method for predicting the next maximum of the sunspot behaviour. This will obviously vary depending on the phase of the cycle when the prediction is made.

In Chapter 6, an extensive investigation was carried out into the possibility of iterating the neural network predictions into the future, using the network's own predictions as future input values. The accuracy of this was found to degrade rapidly if the network was iterated more than once. In particular free-running iterations quickly settled into a periodic pattern regardless of phase when the iteration was started. Although iterating more than once resulted in a rapid loss of accuracy, it was found that one iteration of all the outputs of a network generally provided comparable predictions to larger networks. For example, feeding all 6 outputs of a 36-27-6 net back in as the last 6 inputs provided a 12 ahead prediction which was actually better than the best network with 12 outputs. Iterating again to 18 months ahead, however, was considerably worse than the original 18 month prediction. The hope of using iterated predictions for long term sunspot maximum estimation was found to be groundless.

Two other variations which were attempted were multi-step prediction and a combined neural network and McNish and Lincoln approach.The former was found to be comparable in accuracy with the original predictions although never quite better. The latter method, as described in Section 6.3, never threatened to achieve the same level of prediction accuracy as most other variations.

Finally the method first suggested by Koons and Gorney (1990) was investigated to provide sunspot

maximum predictions. The conclusions reached from that section are summarised in detail in Section 6.5. The basic result is that the whole issue of maximum prediction remains open although this methodology is at least as accurate as other current techniques, although perhaps not superior. The actual predictions of cycle 22 were easily within ±10% for both time and size of maximum. Reasons for why this cycle might have been easier to predict, however, were discussed although these are taken together with the fact that the cycle was different in that it peaked much earlier than all other solar cycles to date. Further investigations on this method have been initiated.
In overall conclusion, the best neural network predictions were produced from the networks with multiple inputs and outputs. While some variations on this were comparable in accuracy, none were better. As stated at the start of this chapter, in the majority of cases these predictions were much better than the benchmark of the McNish and Lincoln approach. In terms of sunspot maximum prediction, the method of Koons and Gorney, multiple output neural networks and the ML method were all tested and found to be comparable in many respects, depending on the form of the individual cycles. It remains now to introduce some areas where extended ideas may prove useful.

7.3 Future Work

There are several aspects which remain to be discussed in this final complete section, which also allows for some more speculative ideas to be proposed. One outstanding area from the work completed is the necessity to examine the delay effect in more detail, in the hope of limiting its effect on neural network predictions. The vast majority of the discussion in this thesis has been concentrated on what can be classed the medium to long term variability of the Sun and has left out the variations which are observed over the space of a few days, including analysis of the more violent solar phenomena such as solar flares. Another posssible variation on the theme so far is to include more than one activity index as an input parameter to a neural network. Ideas for this are suggested briefly. Lastly is a variation on the standard feed-forward neural network architecture called *recurrency* which potentially may offer further rewards when analysing some or all of the problems so far discussed.

7.3.1 Overcoming the Delay Effect

The question of overcoming a delay effect when attempting to predict the future values of any time series is not new. In particular it is at a maximum (or minimum) that one possible reason for this effect is most easily visualised. Since it is at the maxima that the time series switches from increasing to decreasing (or the time derivative changes from positive to negative), so the inputs of the network may indicate a gradual increase in the data encouraging the network to respond

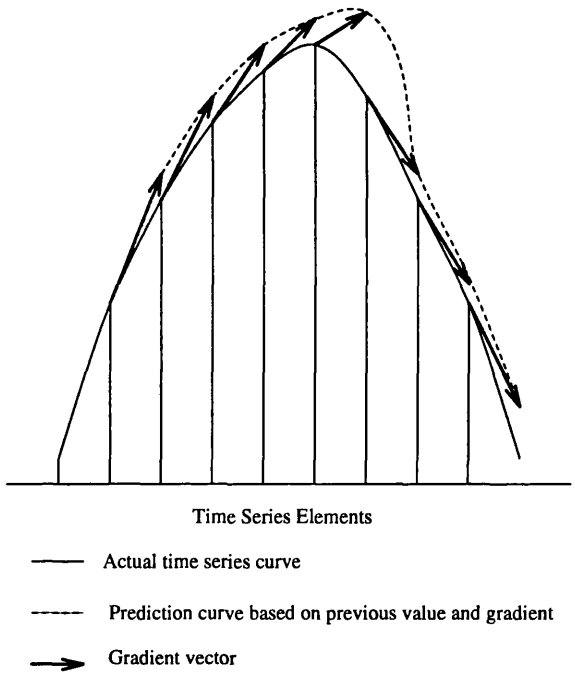


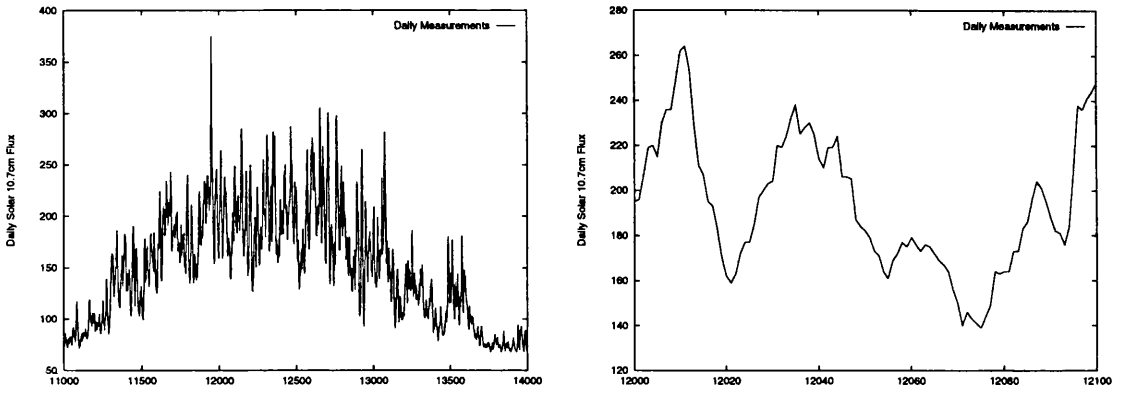
Figure 7.1: Using a simple gradient extrapolating prediction scheme the delay in peak prediction is demonstrated. Although predicting with a NN is very different in its method, this illustrates the same problems faced by all prediction techniques.

with increasing values as outputs. It is only as the network is presented with decreasing inputs that it definitely detects the presence of the maximum. This, albeit oversimplified, theory could explain why the delay effect is most prominent in the more complex, rapidly varying data (see Figure 7.1). Obviously, the more complex the time series or the shorter time resolution used, the more fluctuations in direction exist in the data making the possibility of inclusion of a delay in the predictions more likely. This is evidenced with the comparison between the smoothed and unsmoothed monthly sunspot number. Hence it would be worthwhile to carry out a study to determine whether the delay timescale is constant throughout the prediction set or whether it shows a variation which is dependent on the phase of the solar cycle.

One suggested remedy for the delay effect is to include the gradient of the slope between adjacent points as another input. This would require re-formulating the back-propagation learning algorithm to include some *penalty* for delay. Some of the initial work on including this data as input is currently in progress. An alternative approach which is also currently under study (Conway, 1994) is the use of Genetic Algorithms. The possible advantage in this approach is the relative ease with which a new error term can be included into training. Training neural networks using genetic algorithms has so far been comparable in both speed and prediction accuracy to the standard back-propagation learning mechanism.

7.3.2 Recurrent Network Architectures

The whole of this thesis has concentrated on the application of multi-layer feed-forward neural network architectures learning through back-propagation of errors. Even within the introductory chapter on neural network methods, only the above type of architectures were discussed in detail with, in the main, references provided for some of the alternate theories. At this point the idea of *recurrent* networks is introduced. These vary from those discussed previously in that connections are now allowed both ways between a pair of units and even from a unit to itself. There are many ways of making a network recurrent and it is impossible to make any significant impact in describing all the theories here. As an introduction to the ideas involved, the reader is directed to chapter 7 of Hertz et al. (1991a). The only example of recurrent networks which are quoted here are those which have been proposed to learn time sequences. Most of the theory behind these ideas is not as relevant to the work of this thesis as might be thought. This is because the theory really concerns sequence recognition and not necessarily sequence reproduction or crucially *extension*. Partially recurrent networks are mainly feed-forward but allow for feedback connections to be included. These potentially allow the network to remember cues from the past but does not necessarily complicate the training; in most cases back-propagation can still be applied. The idea of feedback connections is slightly analogous to the idea of a controller in a control system and this has potential application to keep track of gradient terms in the time series. Hence the idea that a suitable trained partially recurrent network may incorporate some features which would be advantageous in overcoming the problem of the delay effect (as described in the previous section). One example of a recurrent architecture suggested by Elman (1990) is mentioned. In this the input layer is divided into two parts, the true inputs and other units called *context* units which receive feedback from the hidden units of the network. In fact these units simply hold copies of the hidden activation values from the previous time step. All the connections are feed-forward and thus modifiable using back-propagation. In this respect the context units are treated as standard inputs. Some initial research was carried out using this architecture but no conclusions were reached. Problems which are encountered when considering the time series prediction desired in this thesis is the necessity to present all the data (including the prediction set) in chronological order. This is obvious from the fact that the point of the context units is to hold cues from the past. Thus having successfully removed the explicit time-dependence from the earlier formulation of the problem, this theory would require the reintroduction of a time parameter. Nevertheless the idea of recurrent networks is suggested as an area where the theoretical aspect of neural network research may be applied to the prediction of solar activity, perhaps with the goal of overcoming the delay effect.



(a) Cycle 21

(b) Sample 100 days

Figure 7.2: The daily behaviour of the solar 10.7cm flux during solar cycle 21 from 1976 to 1986. The figure on the right zooms in on a sample 100 days chosen at random from a segment near the maximum of the cycle.

7.3.3 Short-term Solar Variations

While the medium to long term variations (ie months to years) of solar activity are important to the solar physics community and also in the planning and scheduling of space missions, the daily variations in the solar flux which are observed and measured are crucial in determining the final few orbits of a re-entering satellite (Lobochev et al., 1992). The behaviour of the daily solar flux is extremely complex as witnessed in Figure 7.2(a). One possible way in which analysis of predicting this problem may proceed is through splitting up this data into very many short segments, such as Figure 7.2(b). In this way the network is trained on very many examples showing what the next days solar flux value was, based on, say, the previous weeks' behaviour. Allowing for sufficient computing power to build up a large enough database may allow some degree of accuracy to obtained in this problem. The fact that the network sizes are restricted to being relatively small allows for the inclusion of many more input-output patterns.

Descending further to even shorter timescales is the possibility of analysing solar events which occur in the space of seconds or minutes. The Solar Maximum Mission, launched in 1980, carried the Hard X-Ray Burst Spectrometer (HXRBS) which made many recordings of HXRB events during the lifetime of the mission. A Hard X-Ray Burst is an increase in the HXR flux (about 20keV - 200keV in the electromagnetic spectrum) over a timescale of typically a few minutes and is usually associated with a solar flare event. Such a high energy burst could have implications for the functioning of space-borne instrumentation and for the safety of astronauts. A typical event as recorded by HXRBS (after data reduction) is pictured in Figure 7.3. A HXRB event usually consists of an initial steep rise in flux which reaches a peak and then decays either just as steeply

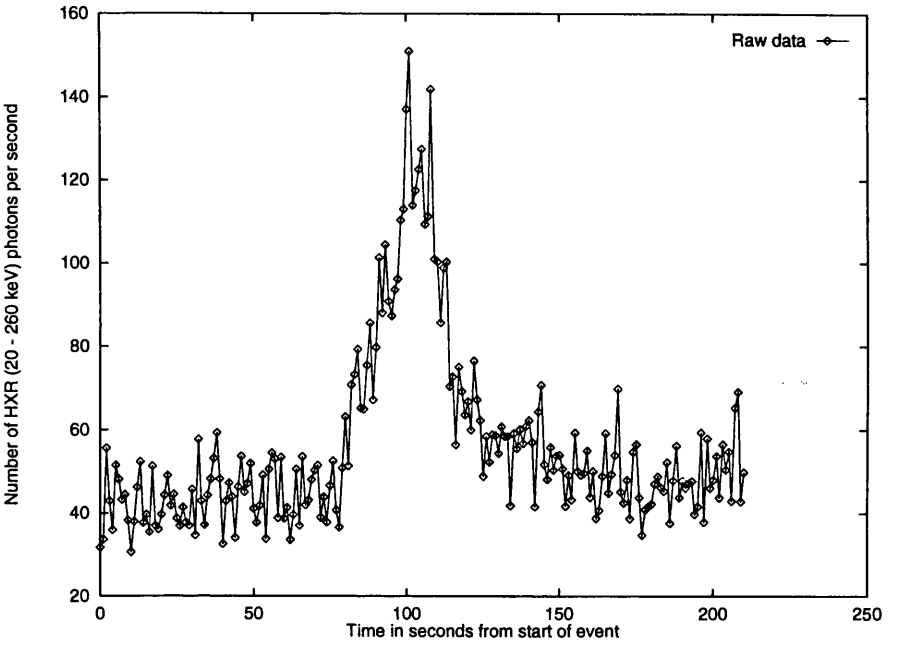


Figure 7.3: A typical raw data HXRB event.

or more slowly with an exponential form. The type of decay is determined by the conditions in the flare region on the Sun. The sharp rise is attributed to the acceleration of electrons in areas of high magnetic field and the decay is representative of the electrons returning to their original (and relatively small) velocities. The mechanism by which the electrons are accelerated is not well understood, inviting the use of neural networks to predict the outcome of such events.

An initial analysis of these events has been considered (Conway, 1993). The treatment using neural networks departs from a strict time series analysis as instead the problem is to recognise the evolution of a burst given the initial behaviour. This is directly analogous to the research outlined in Section 6.4.1 where the evolution of a solar cycle is sought given the onset behaviour. Although the timescales are entirely different, the formulation of the problem removes any explicit time dependence from the inputs.

This work is also relevant to that proposed above where the question was to predict the daily fluctuations in solar flux values. The latter could be considered in terms of a rapid rise (or fall) in solar 10.7cm flux from a background level over the space of a few days, similar to the sudden increase in HXR flux. The future potential of this research remains to be seen depending on the development of the initial work so far attempted.

7.3.4 Multiple-Input Parameters

One final extension of the research outlined so far would be to develop the time series analysis approach from learning on one activity index to the use of two or more indices concurrently. This approach is particularly evident in previous work where the prediction of events is required, for example solar flare forecasting (Shaw, 1993a and Aso and Ogawa, 1993) or predicting the occurrence of geomagnetic storms (Lundstedt, 1992). In the former several properties of sunspot groups, for example magnetic class or Zurich classification, are included as inputs to the network. The work of Lundstedt includes various parameters describing the properties of X-ray flares or CME's which are observed on the Sun.

The precursor methods of predicting the upcoming solar maximum have in some cases included information from geomagnetic behaviour (for example Ohl, 1976 or Schatten et al., 1978). This leads to the suggestion that doubling up some of the indices may provide further additional information which would allow more accurate predictions of the sunspot maximum, for example.

Multiple input parameters are most likely to succeed in the situation where Input A, B $\rightarrow$ Output C, where C is determined by the factors A *and* B. The situation of using A, B to predict A or B is not likely to be of benefit since if the input parameters are well correlated then there is a redundancy in the information while a poor correlation is obviously of little use either. The obvious potential problem with this would be the inclusion of even more data which requires to be analysed and included in network training. In this respect the techniques to optimise network training, through adapting the learning rate at each training step, monitoring the generalisation capabilities of the networks and perhaps including network pruning methods, may provide solutions to the ever increasing amount of data to be handled.

7.4 The Way Forward

As a final summary the work carried out in my research is restated here.

- Development of software to implement the back-propagation of errors learning algorithm in feed-forward neural networks.
- Comprehensive tests of various architectures and prediction styles for the solar activity indices and investigation of ways to monitor the training for optimal generalisation ability.
- The first application of neural networks to predicting the unsmoothed monthly sunspot number and the geomagnetic *aa* index.
- Discovery of a delay effect in neural network predictions.
- Comparison of the prediction accuracies achieved with the McNish and Lincoln method. The former were found to be at least as accurate in every case and considerably more accurate in most cases.

• Extension of previous investigations into neural network prediction of sunspot maximum.

In terms of the future development of neural network techniques, the results here have shown them to be accurate enough to justify their use as methods of prediction and analysis in this, and potentially many other, aspects of astronomy and astrophysics. Currently these methods exist predominantly as tools for statistical and pattern recognition purposes and do not aim to provide additional understanding of the underlying physics involved in the problem. In future it may be possible to reconstruct some aspects of the underlying dynamics of a problem from consideration of network training although this will require more collaboration between network theorists and those who use them in practical applications. More likely is the case that neural networks will remain, in the words of John Denker, the second best way to approach a problem. The best of course is to understand the physics of a problem and express this mathematically. So long as this cannot be done, however, neural networks may provide a more accurate analysis of the data obtained. This is the current situation in understanding solar activity. What I have shown is that until the dynamo action of the Sun is better understood, neural networks will be a very useful tool in analysing and predicting the data observed now.

Bibliography

Anderson, J. A.: 1968, *Kybernetik* **5**, 113

Anderson, J. A.: 1970, *Mathematical Biosciences* **8**, 137

Anderson, J. A. and Mozer, M. C.: 1981, in G. E. Hinton and J. A. Anderson (eds.), *Parallel Models of Associative Memory*, pp 213–236, Erlbaum, Hillsdale

Anderson, J. A. and Rosenfeld, E.: 1988, *Neurocomputing: Foundations of Research*, MIT Press

Aso, T. and Ogawa, T.: 1993, in J. A. Joselyn, H. Lundstedt, and J. Trolinger (eds.), *Proceedings of the International Workshop on Artificial Intelligence Applications in Solar-Terrestrial Physics*, pp 77–82, NOAA Space Environment Laboratory

Babcock, H. D.: 1959, *The Astrophysical Journal* **130**, 364

Babcock, H. D. and Babcock, H. W.: 1955, *The Astrophysical Journal* **121**, 349

Block, H. D.: 1962, *Reviews of Modern Physics* **34**, 123, Reprinted in Anderson and Rosenfeld (1988)

Boteler, D. H.: 1993, in J. A. Joselyn, H. Lundstedt, and J. Trolinger (eds.), *Proceedings of the International Workshop on Artificial Intelligence Applications in Solar-Terrestrial Physics*, pp 183–191, NOAA Space Environment Laboratory

Bravo, S. and Rivera, A. L.: 1994, *The Solar Causes of Major Geomagnetic Storms*, preprint

Bray, R. J. and Loughhead, R. E.: 1965, *Sunspots*, Wiley, New York

Brown, G.: 1988, *Nature* **333**, 121

Brown, G. M. and Simon, P. A.: 1986, in P. Simon, G. Heckman, and M. Shea (eds.), *Solar-Terrestrial Predictions: Proceedings of a Workshop at Meudon, France June 18-22, 1984*, pp 1–7, NOAA, Boulder, CO

Bryson, A. E. and Ho, Y. C.: 1969, *Applied Optimal Control*, Blaisdell, New York

Caianiello, E. R.: 1961, *Journal of Theoretical Biology* **1**, 204

Calvo, R. A., Ceccatto, H. A., and Piacentini, R. D.: 1993, *Neural network prediction of solar activity*, preprint communication

Cater, J. P.: 1987, in M. Caudill and C. Butler (eds.), *IEEE First International Conference on Neural Networks*, Vol. II, pp 645–651, IEEE, New York

Chapman, G. A.: 1984, *Nature* **308**, 252

Conway, A.: 1994, Private communication

Conway, A. J.: 1993, in J. A. Joselyn, H. Lundstedt, and J. Trolinger (eds.), *Proceedings of the International Workshop on Artificial Intelligence Applications in Solar-Terrestrial Physics*, pp 31–36, NOAA Space Environment Laboratory

Cowan, J. D. and Sharp, D. H.: 1988a, *Quarterly Reviews of Biophysics* **21**, 365

Cowan, J. D. and Sharp, D. H.: 1988b, *Dædalus, Proceedings of the American Academy of Arts and Sciences* **117**, 85

Cybenko, G.: 1988, *Continuous Valued Neural Networks with Two Hidden Layers Are Sufficient*, Technical report, Department of Computer Science, Tufts University, Medford, MA

Cybenko, G.: 1989, *Mathematics of Control, Signal and Systems* **2**, 303

Douglass, A. E.: 1919, *Carnegie Institute of Washington Publication 289* **1**, 102, Washington, D.C.

Dungey, J. W.: 1961, *Physics Review Letters* **6**, 47

Eddy, J. A.: 1976, *Science* **192**, 1189

Eddy, J. A.: 1977, in O. White (ed.), *The Solar Output and its Variations*, pp 51–71, Colorado Associated University Press, Boulder

Eddy, J. A., Gilman, P. A., and Trotter, D. E.: 1976, *Solar Physics*

Elman, J. L.: 1990, *Cognitive Science* **14**, 179

Euler, H. and Holland, R.: 1982, in *Proceedings of a Workshop on Satellite Drag, Boulder*, pp 151–161

Feynman, J.: 1982, *Journal of Geophysical Research* **87(A8)**, 6153

Feynman, J. and Crooker, N. U.: 1978, *Nature* **275**, 626

Feynman, J. and Gu, X. Y.: 1986, *Reviews of Geophysics* **24**, 650

Foukal, P., Fowler, L. A., and Livshits, M.: 1983, *The Astrophysiacl Journal* **267**, 863

Foukal, P. and Lean, J.: 1986, *The Astrophysical Journal* **302**, 826

Gleissberg, W.: 1958, *Journal of the British Astronomical Association* **68**, 148

Goldberg, D. E.: 1989, *Genetic Algorithms in Search, Optimisation and Machine Learning*, Addison-Wesley

Gonzalez, W. and Tsurutani, B. T.: 1987, *Planetary and Space Science* **35**, 1101

Gorney, D. J.: 1989, *Journal of Spacecraft and Rockets* **26**, 428

Gorney, D. J.: 1990, *Reviews of Geophysics* **28**, 315

Gosling, J. T., McComas, D. J., Phillips, J. L., and Bame, S. J.: 1991, *Journal of Geophysical Research* **96**, 7831

Grossberg, S.: 1967, *Proceedings of the National Academy of Sciences, USA* **58**, 1329

Grossberg, S.: 1968a, *Proceedings of the National Academy of Sciences, USA* **59**, 368

Grossberg, S.: 1968b, *Proceedings of the National Academy of Sciences, USA* **60**, 758

Grossberg, S.: 1969, *Journal of Mathematical Psychology* **6**, 209

Grossberg, S.: 1972, *Kybernetik* **10**, 49

Grossberg, S.: 1976a, *Biological Cybernetics* **23**, 121, Reprinted in Anderson and Rosenfeld (1988)

Grossberg, S.: 1976b, *Biological Cybernetics* **23**, 187

Grossberg, S.: 1980, *Psychological Review* **87**, 1, Reprinted in Anderson and Rosenfeld (1988)

Grossberg, S.: 1987a, *The Adaptive Brain*, Elsevier, Amsterdam

Grossberg, S.: 1987b, *Cognitive Science* **11**, 23

Hale, G. E.: 1912, *Publications of the Astronomical Society of the Pacific* **24**, 223

Hale. G. E., Hellerman, F., Nicholson, S. B., and Joy, A. H.: 1919, *The Astrophysical Journal* **49**, 153

Hale, G. E. and Nicholson, S. B.: 1925, *The Astrophysical Journal* **62**, 270

Hassibi, B. and Stork, D. G.: 1992, in S. Hansen, J. Cowan, and C. Giles (eds.), *Proceedings of the Neural Information Processing Systems-5*, Morgan-Kaufmann

Hebb, D. O.: 1949, *The Organization of Behavior*, Wiley, New York, Partially reprinted in Anderson and Rosenfeld (1988)

Hedin, A. E. and Mayr, H. G.: 1987, *Journal of Geophysical Research* **92**, 869

Hertz, J.: 1993a, Private communication

Hertz, J.: 1993b, in J. A. Joselyn, H. Lundstedt, and J. Trolinger (eds.), *Proceedings of the International Workshop on Artificial Intelligence Applications in Solar-Terrestrial Physics*, pp 55–64, NOAA Space Environment Laboratory

Hertz, J., Krogh, A., and Palmer, R. G.: 1991a, *Introduction to the Theory of Neural Computation*, Addison-Wesley

Hertz, J. A., Krogh, A., and Palmer, R. G.: 1991b, *Introduction to the Theory of Neural Computation*, Chapt. 6.2, pp 120–130, Addison-Wesley

Hirman, J.: 1989, NOAA Space Environment Laboratory, see Withbroe (1989) for details

Hirman, J. W., Heckman, G. R., Greer, M. S., and Smith, J. B.: 1988, *EOS Transactions, American Geophysical Union* **69**, 962

Hirshberg, J.: 1973, *Astrophysics and Space Science* **20**, 473

Holland, R. L. and Vaughan, W. W.: 1984, *Journal of Geophysical Research* **89**, 11

Hopfield, J. J.: 1982, *Proceedings of the National Academy of Sciences, USA* **79**, 2554, Reprinted in Anderson and Rosenfeld (1988)

Hoyt, D. V. and Kyle, H. L.: 1990, in K. Schatten, A. Arking, and R. Schiffer (eds.), *Proceedings of Conference on the Climatic Impact of Solar Variability*, pp 293–300, NASA, GSFC

Hudson, H. S., Silva, S., Woodard, M., and Willson, R. C.: 1982, *Solar Physics* **76**, 211

Hundhausen, A. J.: 1972, *Coronal Expansion and Solar Wind*, Springer-Verlag, Berlin

Jacobs, R. A.: 1988, *Neural Networks* **1**, 295

Jenkins, R.: 1992, Private Communication

Johnson, M. D. and Adorf, H. M.: 1992, *Computers Ops. Res.* **19(3/4)**, 209

Joselyn, J. A. and McIntosh, P.: 1981, *Journal of Geophysical Research* **86**, 4555

Kane, R. P.: 1987, *Solar Physics* **108**, 415

Kerridge, D., Carlaw, V., and Beamish, D.: 1989, *Development and Testing of Computer Algorithms for Solar and Geomagnetic Activity Forecasting*, European Space Agency Contract Report WM/89/22c, British Geological Society

Kohonen, T.: 1974, in *IEEE Transactions on Computers*, Vol. C-23, pp 444–445

Kohonen, T.: 1982, *Biological Cybernetics* **43**, 59, Reprinted in Anderson and Rosenfeld (1988)

Kohonen, T.: 1989, *Self-Organization and Associative Memory*, Springer-Verlag, Berlin, 3rd edition

Koons, H. C. and Gorney, D. J.: 1990, *Eos Transactions, AGU* **71**, 677

Koons, H. C. and Gorney, D. J.: 1991, *Journal of Geophysical Research* **96(A4)**, 5549

Kunches, J. M., Heckman, G. R., Hildner, E., and Suess, S. T.: 1991, *Solar Radiation Forecasting and Research to Support the Space Exploration Initiative*, Technical report, NOAA Space Environment Laboratory

Lantos, P. and Simon, P.: 1987, in *Proceedings of the 8th ESA Symposium on European Rocket and Balloon Programmes and Related Research*, pp 451–453, European Space Agency, Noordwijk, SP-270

Lapedes, A. and Farber, R.: 1987, *Nonlinear Signal Processing Using Neural Networks: Prediction and System Modelling*, Technical Report LA-UR-87-2662, Los Alamos National Laboratory, Los Alamos, NM

Lawrence, J. K. and Chapman, G. A.: 1990, *The Astrophysical Journal* **361**, 709

Le Cun, Y.: 1985, in *Cognitiva 85: A la Frontière de l'Intelligence Artificielle des Sciences de la Connaissance des Neurosciences*, pp 599–604, CESTA, Paris

Le Cun, Y., Denker, J. S., and Solla, S. A.: 1990, in D. Touretzky (ed.), *Proceedings of the Neural Information Processing Systems-2*, pp 598–605, Morgan-Kaufmann

Lean, J.: 1991, *Reviews of Geophysics* **29**, 505

Lindberg, J. P.: 1989, *Solar Activity Inputs for Upper Atmosphere Models Used in Programs to Estimate Spacecraft Orbital Lifetime*, Technical report, NASA Marshall Space Flight Center, details from Withbroe (1989)

Little, W. A.: 1974, *Mathematical Biosciences* **19**, 101

Little, W. A. and Shaw, G. L.: 1978, *Mathematical Biosciences* **39**, 281

Lloyd-Hart, M., Wizinowich, P., McLeod, B., Wittman, D., Colucci, D., Dekany, R., McCarthy, D., Angel, J. R. P., and Sandler, D.: 1992, *The Astrophysical Journal* **390**, L41

Lobochev, V., Pochukaev, V., Ivanov, N., Dzesov, R., Bidenko, N., Yastrebov, V., Glotov, V.,

Zhukov, V., Korsakov, V., Kiryanov, V., Alekseyev, A., and Motsulyov, B.: 1992, *ESA Journal* **16(2)**, 209

Lonnblad, L., Peterson, C., and Rögnvaldsson, T.: 1990, *Physical Review Letters* **65(11)**, 1321

Lundstedt, H.: 1992, *Planetary and Space Science* **40(4)**, 457

Lundstedt, H.: 1993, in J. A. Joselyn, H. Lundstedt, and J. Trolinger (eds.), *Proceedings of the International Workshop on Artificial Intelligence Applications in Solar-Terrestrial Physics*, pp 199–204, NOAA Space Environment Laboratory

Mandelbrot, B. B.: 1982, *The Fractal Geometry of Nature*, San Francisco: Freeman

Marr, D.: 1969, *Journal of Physiology (London)* **202**, 437

Marr, D.: 1970, *Proceedings of the Royal Society of London B* **176**, 161

Marr, D.: 1971, *Philosophical Transactions of the Royal Society of London B* **262**, 23

Maunder, E. W.: 1890, *Monthly Notices of the Royal Astronomical Society*

Maunder, E. W.: 1922, *Journal of the British Astronomical Association* **32**, 140

Mayaud, P. N.: 1975, *Journal of Geophysical Research* **80(1)**, 111

McClelland, J. L. and Rumelhart, D. E.: 1981, *Psychological Review* **88**, 375, Reprinted in Anderson and Rosenfeld (1988)

McCulloch, W. S. and Pitts, W.: 1943, *Bulletin of Mathematical Biophysics* **9**, 127, Reprinted in Anderson and Rosenfeld (1988)

McNish, A. G. and Lincoln, J. V.: 1949, *Transactions of the American Geophysical Journal* **30**, 673

Miller, A. S.: 1993, *Vistas in Astronomy* **36(2)**, 141

Minsky, M. and Papert, S.: 1969, *Perceptrons*, MIT Press, Introduction Reprinted in Anderson and Rosenfeld (1988)

Montana, D. J. and Davis, L.: 1989, in N. S. Sridharan (ed.), *Eleventh International Joint Conference on Artifiial Intelligence*, pp 762–767, Morgan-Kaufmann, San Mateo

Mugellesi, R. and Kerridge, D. J.: 1991, *ESA Journal* **15(2)**, 123

Müller, B. and Reinhardt, J.: 1991, *Neural Networks: An Introduction*, Springer-Verlag

Mundt, M. D., Maguire II, W. B., and Chase, R. R. P.: 1991, *Journal of Geophysical Research* **96(A2)**, 1705

Ohl, A. I.: 1976, *Soln. Dannye* **9**, 73, (in Russian)

Packard, N. H., Crutchfield, J. P., Farmer, J. D., and Shaw, R. S.: 1980, *Physical Review Letters* **45**, 712

Parker, D. B.: 1985, *Learning Logic*, Technical Report TR-47, Centre for Computational Research in Economics and Management Science, Massachusetts Institute of Technology, Cambridge, MA.

Pomerlau, D. A.: 1989, in D. S. Touretzky (ed.), *Advances in Neural Information Processing*

Systems I (Denver 1988), pp 305–313, Morgan Kaufmann, San Mateo

Rosenblatt, F.: 1958, *Psychological Review* **65**, 386, Reprinted in Anderson and Rosenfeld (1988)

Rosenblatt, F.: 1962, *Principles of Neurodynamics*, Spartan, Washington D.C

Rumelhart, D. E., Hinton, G. E., and Williams, R. J.: 1986a, in D. E. Rumelhart and J. L. McClelland (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1, Chapt. 8, pp 318–362, MIT Press, Reprinted in Anderson and Rosenfeld (1988)

Rumelhart, D. E., Hinton, G. E., and Williams, R. J.: 1986b, *Nature* **323**, 533, Reprinted in Anderson and Rosenfeld (1988)

Rumelhart, D. E. and McClelland, J. L.: 1982, *Psychological Review* **89**, 60

Rumelhart, D. E., McClelland, J. L., and the PDP Research Group: 1986c, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition (2 vols)*, MIT Press

Sargent, H. H.: 1978, in *Proceedings of the 28th IEEE Vehicular Technical Conference*, pp 490–496, Denver

Schatten, K. H., Scherrer, P. H., Svalgaard, L., and Wilcox, J. M.: 1978, *Geophysics Research Letters* **5**, 411

Schatten, K. H. and Sofia, S.: 1987, *Geophysical Research Letters* **14**, 632

Schwabe, H.: 1843, *Astronomische Nacrichten* **21**, 233

Sejnowski, T. J. and Rosenberg, C. R.: 1987, *Complex Systems* **1**, 145

Shaw, D.: 1993a, in J. Joselyn, H. Lundstedt, and J. Trolinger (eds.), *Proceedings of the International Workshop on Artificial Intelligence Applications in Solar-Terrestrial Physics*, pp 71–76, NOAA Space Environment Laboratory

Shaw, D.: 1993b, *Ph.D. thesis*, University of Colorado, Boulder

Sheeley Jr., N. R., Howard, R. A., Koomen, M. J., Michels, D. J., Schwenn, R., Mühlhäuser, K. H., and Rosenbauer, H.: 1985, *Journal of Geophysical Research* **90**, 163

Smart, D. F. and Shea, M. A.: 1989, *Journal of Spacecraft and Rockets* **26**, 403

Sofia, S., Oster, L., and Schatten, K.: 1982, *Solar Physics* **80**, 87

Solla, S. A., Levin, E., and Fleischer, M.: 1988, *Complex Systems* **2**, 625

Sporer, F. W. G.: 1887, *Astr. Ges. Vjschr. Leipzig* **22**, 323

Sporer, F. W. G.: 1889, *Bulletin of the Royal Astronomical Society* **6**, 60

Storrie-Lombardi, M. C., Lahav, O., Storrie-Lombardi, L., and Sodr, L.: 1992, *Monthly Notices of the Royal Astronomical Society* **259**, 8

Takens, F.: 1981, in D. Rand and L.-S. Young (eds.), *Detecting Strange Attractors In Turbulance*, Vol. 898 of *Lecture Notes in Mathematics*, pp 366–381, Springer-Verlag

Taylor, W. K.: 1956, in C. Cherry (ed.), *Information Theory*, pp 314–328, Butterworths, London 1985

Thompson, R. J.: 1987, in *Preliminary Report and Forecast of Solar-Geophysical Data*, p. 12,

National Oceanic and Atmospheric Administration, Space Environment Research Lab., Boulder, CO

Viljanen, A.: 1993, in J. Joselyn, H. Lundstedt, and J. Trolinger (eds.), *Proceedings of the International Workshop on Artificial Intelligence Applications in Solar-Terrestrial Physics*, pp 193–198, NOAA Space Environment Laboratory

Vogl, T. P., Mangis, J. K., Rigler, A. K., Zink, W. T., and Alkon, D. L.: 1988, *Biological Cybernetics* **59**, 257

Waldmeier, M.: 1961, *The Sunspot-Activity in the Years 1610-1960*, Schultess and Co., Zurich

Walterscheid, R. L.: 1989, *Journal of Spacecraft and Rockets* **26(6)**, 439

Weigend, A. S., Huberman, B. A., and Rumelhart, D. E.: 1992, in M. Casdagli and S. Eubank (eds.), *Proceedings of the Workshop on Nonlinear Modeling and Forecasting*, pp 395–432, SFI Studies in the Sciences of Complexity, Addison-Wesley

Werbos, P.: 1974, *Ph.D. thesis*, Harvard University

White, H.: 1990, *Neural Networks* **3**, 535

Widrow, B. and Hoff, M. E.: 1960, in *1960 IRE WESCON Convention Record*, No. 4, pp 96–104, IRE, New York, Reprinted in Anderson and Rosenfeld (1988)

Williams, K. E.: 1991, *Johns Hopkins APL Technical Digest* 12(4)

Wilson, R. M.: 1984, *A Comparative Look at Sunspot Cycles*, Technical Report 2325, NASA

Wilson, R. M.: 1988a, *Nature* **335**, 773

Wilson, R. M.: 1988b, *Solar Physics* **115**, 397

Wilson, R. M.: 1988c, *Journal of Geophysical Research* **93**, 10011

Wilson, R. M., Reichman, E. J., and Teuber, D. L.: 1986, in P. A. Simon, G. Heckman, and M. A. Shea (eds.), *Solar-Terrestrial Predictions Workshop*, pp 26–34

Wintoft, P. and Lundstedt, H.: 1993, in J. A. Joselyn, H. Lundstedt, and J. Trolinger (eds.), *Proceedings of the International Workshop on Artificial Intelligence Applications in Solar-Terrestrial Physics*, pp 37–44, NOAA Space Environment Laboratory

Withbroe, G. L.: 1989, *Journal of Spacecraft and Rockets* **26(6)**, 394

Wright, C. S. and McNamara, L. F.: 1983, *Solar Physics* **87**, 401