AM

MET O 3 TECHNICAL NOTE NO 33

A REVIEW OF THE AREAL QUALITY CONTROL OF DAILY CLIMATOLOGICAL DATA

by

R. C. Tabony

This report has not been published. Permission to quote from it should be obtained from the Assistant Director of the Advisory Services Branch of the Meteorological Office.

Advisory Services (Met O 3)
Meteorological Office
London Road
Bracknell
Berkshire

December 1985

Contents

Summary

The principal component approach currently used for the areal quality control of temperature and sunshine is investigated with a view to optimising the implementation of the technique. A near neighbour scheme was also written to act as a benchmark against which the performance of principal components might be assessed. It was found that both approaches provided a similar standard of quality control. The potential of the principal component technique to obtain improved results was impaired by its sensitivity to errors and skewed distributions. Nevertheless it gained over near neighbours as the data became less well correlated, and hence was able to produce the better estimates at the more isolated stations. The adverse effects of a non-normal distribution also degraded the near neighbour scheme by reducing the efficiency of its flagging routines. The latter approach, however, was found to perform better for new stations for which pre-calculated components were not available. The current implementation of the principal component technique was found to be close to being optimal.

1. Introduction

The quality control of meteorological observations plays an important role in the establishment of a climatological data base. The large volume of data involved requires that the procedures be automated, and the checks which can be devised fall into 5 main classes:-

(i) climatological range checks.

(ii) consistency between elements at each observation.

(iii) consistency between values of a given element at successive observations.

(iv) consistency in space at a given point in time.

(v) homogeneity checks on site and instrumental changes.

Consistency in time is very useful for observations made so frequently that only small changes in the element are expected between successive observations; this applies to hourly 'synoptic'observations of most elements and daily recordings of soil temperature. Spatial comparisons become necessary for 'climatological' observations made only once per day.

This paper is mainly concerned with spatial consistences checks, especally those relating to daily observations of minimum temperature and sunshine. For wind and rain sophisticated near neighbour techniques are well established whereas for temperature and sunshine a simple near neighbour approach (Bryant, 1979) was replaced by a principal component method (Spackman and Singleton, 1982). In making this innovation a number of arbitrary decisions were made and a major aim of this work is to explore the scope for optimisation of this new approach. A sophisticated near neighbour scheme was also written to act as a benchmark against which the performance of the principal component technique might be assessed. The discussion is restricted to the use of the conventional station network although it is recognised that remote sensing may have a future role to play.

2. The data used and some of its characteristics

Temperature data was extracted for every third day from 1973 to 1982, a period which included the years 1973 to 77 used by Spackman (1980) and which enabled his procedures to be reproduced and tested on independent data. The serial correlation between daily recordings of minimum temperature (0.6) is such that independent values can be secured by using observations from only every third day. For sunshine, however, the serial correlation of 0.2 made it worthwhile using daily observations, and the period chosen was 1979 to 83.

Principal component analysis should only be applied to a correlation or covariance matrix which is complete and self-consistent, and one of the simplest ways of achieving this is to compute it from a complete set of data. The stations used were therefore limited to those with 90% of data availability, with the remaining values estimated as having the same temperature anomaly or percentage of average sunshine as that obtained from the mean of half a dozen or so neighbouring stations. This reduced the number of stations considered to 373 for temperature and 197 for sunshine, about one half to two-thirds of the full network, and their distributions are shown in fig 1.

One of the most important criteria for the success of spatial quality control is the extent of correlation amongst the station network. The highest interstation correlations for the present network averaged 0.92 for both minimum temperature and sunshine, and their distributions are shown in fig 2. This equality of mean values conceals a fundamental difference in the correlation characteristics of the two elements. Hopkins (1977) shows that for sunshine, the correlation between 2 points close together approaches unity and that thereafter, over the uncomplicated terrain of East Anglia, it falls off at the rate of 0.23 per 100 km. For minimum temperature, the correlation for points close together is 0.93, but thereafter falls off at the rate of only 0.05 per 100 km. These differences are illustrated by the fact that for 8th ranked neighbours, the correlation is 0.87 for minimum temperature and 0.80 for sunshine.

Another distinction between temperature and sunshine is that while the distribution of temperature is reasonably normal, that for sunshine is markedly non-normal, displaying U-shaped characteristics. This is likely to pose problems for all techniques based on the assumption of linearity.

3. The distribution of errors

In order to assess the performance of a given technique for quality control, it is necessary to test it on data which has been contaminated with a realistic distribution of errors. A guide to this was obtained by examining the amendments made by the quality control staff to the climatological returns for 50 to 100 stations in England and Wales during 1982-83. Errors of less than 2 deg C or 2 hours of sun were ignored, and only amendments ascribed by the writer to areal quality control were considered. For minimum temperature this procedure yielded errors with a frequency of about 1 observation in 100 and a magnitude which could be modelled by a gamma distribution with parameters α =2 and β =1.

For sunshine the frequency of errors was 1 in 190 observations but their magnitude varied seasonally in proportion to the maximum possible amount of sun. They could be adequately represented by a truncated normal distribution with a threshold and standard deviation of 17 and 30 per cent of maximum possible sun respectively. The observed and modelled distribution of errors for minimum temperature and sunshine are both displayed in fig 3.

Errors were inserted into the data at regular intervals of 100 or 190 observations by a random number generator with the appropriate distribution. The regular insertion of errors with respect to stations is unrealistic, but is not expected to invalidate the analysis. It is more difficult to assess the results if certain stations are chosen to have more errors than others as the performance of the quality control will then depend on the site characteristics of the stations.

4. Principal component analysis

Introduction

Principal component analysis provides an efficient means of reducing the dimensionality of a problem and is fully described in standard texts (eg Kendall, 1980). In the present content, each component can be regarded as representing a characteristic spatial pattern of observations. The spatial pattern represented by the observations on a given day can then be expressed as a linear combination of patterns represented by the components.

Principal component analysis is now a very popular tool in meteorology and has two main uses - data exploration and data reduction. Most meteorological applications have been concerned with data exploration, in which the rotation of the components to aid interpretation has been widely practised and discussed. (eg Richman, 1986). Quality control applications are essentially concerned with data reduction, however, and so rotation of the components is unnecessary.

The terminology used in principal component analysis is somewhat inadequate and often confusing so the notation adopted in this account is now described. The input to principal component analysis comprises a data array of variables by cases. In the present context, this will be an array of stations by days, and variables can be assigned to either stations or days (referred to as S and T mode analyses respectively). The output takes two forms:-

(i) an array of vectors comprising variables x components. For each component, the vector constitutes a string of numbers, one for each variable, which represents a characteristic distribution of the variables.

(ii) an array of coefficients comprising components x cases, so that, for each component, the coefficients represent the weighting of the component for each case.

Estimates of the original data can be obtained by taking the product of the vectors and the coefficients for each component, and summing over the components:-

$$\text{data} = \sum_{\text{comps}} \text{vectors x coeffs}$$

Successively closer approximation are obtained as the number of components included in the summation is increased; when all the components are used, the original data is reproduced exactly.

In the above description the weight of a component is reflected in the size of the coefficients (the RMS value of the coefficient = $\sqrt{}$eigenvalue). A common transformation is to divide the coefficients, and multiply the vectors, by the square root of the eigenvalue. The standardised coefficients are then known as 'scores' while the transformed vectors are known as 'loadings'. We then have

$$\text{data} = \sum_{\text{comps}} \text{loadings x scores}$$

Computing constraints

The computing time and space required to perform a principal component analysis are proportional to V^3 and V^2 respectively, where V is the number of variables supplied. On the IBM 3081 storage is sufficient to deal with 600 variables (program size 4000K) whereas time constraints set the limit at about 400 variables (time = 5 minutes using the NAG routine F02BBE). Recent improvements in the Met Office's IBM computer installation have increased the storage available rather than reduced the computation time. Thus at the time when Spackman was developing the current system, storage requirements would have limited the number of variables which could be dealt with to about 200.

The climatological network in the UK contains about 600 stations so it is not possible to perform S mode analysis on the full network. Autocorrelation in temperature series causes independent observations to be obtained on about one day in three. By limiting the data supplied to principal component analysis to one day in three, therefore, a years daily data for one station can be represented by 120 observations. Thus a direct T mode analysis could be performed by Spackman on one year of data, but not two or more. In order to overcome this constraint, Spackman devised the following technique to obtain 'averages' of principal components over several years.

(i) The complete set of data (eg 5 years) is divided into segments of a size which can be handled by the available computing facilities (eg 1 year).

(ii) A T mode analysis is performed on each of the segments (years).

(iii) The spatial patterns are represented by the coefficients and those associated with the leading components (eg the first 25) are retained.

(iv) The leading coefficients from each segment are grouped together (to form, say,125 components) and submitted as data to a second principal component analysis (see fig 4).

(v) If the 25 components from each segment were exactly the same, the second principal component analysis will reduce the 125 old components to 25 new ones, each exactly the same as the original 25. The spatial patterns will again be represented by the coefficients.

The benefits of this technique apply when the components obtained from the first analyses are nearly, but not quite, stable. If the components obtained from the segments are very different from one another, then interpretation of the second analysis is uncertain. If the components from the segments are very stable, however, then little is to be gained by performing the second analysis.

The current areal quality control system

The principal components used in the current quality control system were obtained as follows:-

(i) Input consisted of a covariance matrix derived from data expressed in the form of departures from the mean for the day. This meant that long period station averages were not required.

(ii) Variables were assigned to days. This was imposed by computing constraints if data for the whole of the UK was to be supplied at one time.

(iii) Data were obtained from every third day during the period 1973-77 (later extended to 1972-79).

(iv) Principal components were obtained for each year separately and then 'averaged' by performing a second principal component analysis. This step was imposed by computing constraints.

(v) Data were supplied for the complete station network, ie separate analyses were not made for smaller regions.

(vi) Data were supplied for all seasons, ie separate analyses were not made for different seasons.

(vii) The number of components to be used in estimating data was set at 15.

(viii) The 15 leading components were rotated. Although this alters each component, it does not alter their sum. The step was included for purposes of data exploration rather than data estimation.

The assignment of stations to cases means that the spatial patterns correspond to coefficients or scores rather than vectors. The spatial distribution of observations on a given day are therefore regressed against the scores to find the weight, or loading, of each component on the day. Estimates of the observations are made by summing the product of the scores and loadings over the 15 components, and flags are raised if the observations and estimates are not in reasonable agreement.

One sixth of the time of 4 persons is taken up in the manual scrutiny and support of the areal quality control routines.

5. Review of techniques

Spatial interpolation V near neighbours

Climatological quality control involves making an estimate of an element at a station and then comparing it with the observed value. If the element varies smoothly in space then the most obvious way of obtaining an estimate is by spatial interpolation. This condition is unlikely to be satisfied where small-scale topographic effects are important, eg for wind and minimum temperature.

Since quality control does not require estimates to be made at any place, but only where observations are already available, the so-called near neighbour technique can be used to overcome the topographic effects. This involves using past data to identify the most highly correlated stations, developing regressions to estimate the value at the test station, and then combining the estimates by attaching weights according to the correlations. In this technique the regression step may be regarded as 'reducing' observations at neighbouring stations to values appropriate to the site characteristics of the test station. An alternative to the final combination step is therefore spatial interpolation among the 'reduced' observations.

If there is a gradient in a climatological element across an area and the test station is located near the centre of the chosen neighbours then spatial interpolation will provide an estimate close to the mean of the reduced neighbouring values. This will be similar to that provided by the combination step. If the test station is to one side of the network of neighbours then the estimate provided by the combination step may not be very good but then spatial interpolation will involve an element of extrapolation. In general, the neighbour with the greatest weight will be that which is nearest to the test station in spatial interpolation and that which has been most highly correlated in the combination step. So it is not obvious that spatial interpolation offers any general advantages over a simple combination step.

Near Neighbours v Principal Components

The principal component and near neighbour techniques both use relationships apparent during a past period of data to estimate observations at a station from its neighbours. When operating on clean, normally distributed data, principal component analysis is clearly the superior technique. The extremes of minimum temperature observed on radiation nights at frost hollows, for instance, cannot be reproduced by the near neighbour regressions. Principal component analysis, on the other hand, is able to make estimates by attaching a large weighting to the distribution of minima which actually occurred on previous radiation nights. For quality control purposes, however, when the test data contain errors, the advantages of principal components are less. The weighting of a component on a given day is found by matching the pattern of daily observations against that represented by the component, and the errors in the data lead to incorrect assessments of the weightings.

The leading principal components represent large-scale patterns in which most of the stations contribute substantially towards the variance. As higher order components are considered, the number of stations contributing significantly to the pattern decreases - the final components are liable to represent idiosyncracies in the observations at individual stations. There is unlikely to be any serious mis-matching of the true pattern of observations against the leading components, as mistakes in only 1% of observations, for instance, are unlikely to upset estimates of N-S or E-W gradients across the country. The probability of a mis-match is much greater for the higher order components, for which the number of stations contributing substantially to the pattern is much smaller. If a component was determined largely by the behaviour of only 10 stations, for example, then an error in an observation at one of the stations would seriously upset the weighting to be attached to that component.

In a typical near neighbour scheme, the final estimate is a weighted mean of estimates obtained from neighbouring stations and an error at the test station does not impair the quality of the estimate; this depends only on the quality of observations at the neighouring stations. The effect of these can be minimised by scanning the observations, identifying any outliers (which are likely to be errors) and preventing them from contributing to the final estimate. In other words, near neighbour techniques can be made robust to the presence of errors, whereas principal component analysis is sensitive to errors in the data.

6. Estimation by principal components

The difference between an estimate and the true value of an observation is referred to as a residual, and a simple way of assessing the quality of the estimates is to calculate the RMS value of the residuals. In doing this, however, it is necessary to distinguish between correct and erroneous observations. A consequence of the matching of an error-contaminated data set against the components is that it is the erroneous observations, rather than their true values, which are being matched. It follows that the residuals of the erroneous observations are greater than those of the correct observations.

The residuals of correct observations continue to decrease as the number of components is increased. When all the components are used, the input data will be reproduced exactly, even though it is independent of that used to construct the components. The situation is analogous to the fitting of a series of N data points with an N-dimensional polynomial. The use of all components will not result in good error detection, however, since the errors will be as well fitted as the correct observations. As discussed earlier, it is mainly the higher order components which are responsible for the over-fitting. Consequently, as the number of components is increased, the residuals of erroreous observations will decrease and reach a minimum before increasing. In this section, therefore, attention is concentrated on the residuals of erroreous observations.

A thorough testing of the principal component technique involves a consideration of the following points.

(i) S or T - mode analysis

In S-mode analysis the variables are assigned to stations and the cases to days and the spatial patterns are represented by the vectors or loadings, and the corresponding time series by the coefficients or scores. In T-mode analysis the variables are assigned to days and the cases to stations, and the spatial patterns are represented by the coefficients or scores and the time series by the vectors or loadings.

In quality control the object is to match the spatial distribution of observations on one day against the spatial patterns of the components and to find the weighting of each component on the day. In S-mode analysis this can be done analytically, since the weighting of the pattern corresponds to its coefficient, which can be found from the relationship

$$\text{coefficient} = \sum_{\text{variables}} \text{data x vectors}$$

In T-mode analysis, however, the weighting of the pattern corresponds to the vector, and this can only be found by regressing the daily observations against the component coefficients or scores.

Identical results are obtained from S and T-mode analyses. The orthogonal nature of the independent variables in the T-mode regression ensures that it provides a perfect fit, and the spatial patterns associated with the vectors of the S-mode analyses are reproduced by the coefficients of the T-mode analysis.

(ii) Form of input data

Principal component analysis operates on either a correlation or covariance matrix. For temperature and sunshine the covariance matrix is considered more appropriate, but this can be calculated in several ways. One could submit untransformed climatological data but it is helpful to remove the grosser features of the seasonal variation. This can be done by using departures from the long period station mean, referred to as the 'station' anomaly, or departures from a mean over all stations for the day, the 'daily' anomaly. Alternatively, one could use a combination, the 'local' anomaly, based on the difference between a station anomaly and the mean station anomaly on that day. For sunshine one has the additional complication of choice of unit - hours, percent of possible, or percent of average. The choice of input data is therefore any combination of unit and anomaly, although they are not all independent.

The use of local anomalies is associated with a smaller variance of the input data and table 1 shows that, when only a small number of components are used, this advantage is retained. For the number of components commonly required to achieve the smallest residuals, however, this advantage is lost and very similar results are obtained for all types of unit and anomaly. This similarity gives the advantage to the use of hours and the daily anomaly, since this combination does not require the availability of long period station averages.

(iii) Area covered

The leading principal components form a much better representation of events at standard sites in the middle of the area examined, or in dense networks of stations, than those in isolated or unusual sites on the fringe of an area. In the UK, the most isolated sites are also found on the edge of the area examined, namely in the north-western half of Scotland. Better performance may therefore be obtained by examining separately small compact areas of the UK where the density of stations is fairly uniform and where the reasonably circular shape helps to limit edge effects. Accordingly the UK was divided into 10 regions whose boundaries are indicated by the dotted lines in fig 1. The main aim was to avoid long land boundaries and hence minimise edge effects, but in western Scotland the main concern was to provide an area with a uniform distribution of stations.

A comparison of regional and UK analyses is presented in table 2, where it can be seen that the smallest residuals of erroneous observations is reached using 3 components for a regional analysis and more than 24 components for a UK analysis. Although residuals for the majority of observations are smaller for the regional than the UK analysis, the reverse is true for the erroneous observations. Hence it may be concluded that separate regional analyses confer no overall benefit over a single UK analysis.

(iv) Seasonal effects

Characteristic patterns of meteorological elements vary with season so best results may be obtained by restricting input data to a single season. Possible benefits will, however, be offset by the reduced number of days available for analysis. Table 3 compares residuals for individual months (the mean of January, April, July and October) obtained from annual and 3 monthly analyses of sunshine based on 2 years of data. The seasonal analyses are seen to be slightly inferior to those based on the whole year.

(v) Number of years data

The performance of the principal component technique clearly improves with the length of data supplied, but table 4 shows that the effect is small, with RMS residuals of minimum temperature decreasing by only 0.07 deg C as the time base is increased from 1 to 5 years. This improvement will be offset by the need to cope with new stations, for which components will not be available and will need to be estimated. The best number of years to use will therefore depend on the trade-off between the improvement for established stations, the number of new stations opening each year, and the accuracy with which their components can be estimated.

(vi) Use of principal component analysis to 'average' results of previous analyses

Computing constraints prevented Spackman from obtaining components for 5 years directly and so he used a second principal component analysis to 'average' the results of 5 annual analyses. The efficacy of this procedure is investigated in table 5. It shows that the residuals from annual analyses for the years 1973 to 77 are all very similar, and although the spatial patterns have not been examined, it is presumed that at least the leading ones are stable. The last column presents the residuals obtained by subjecting the coefficients obtained from the years 1973 to 77 to a second principal component analysis, and the results may be compared with those in the previous column obtained from a direct analysis made on the period 1973-77. It can be seen that the second principal component analysis has indeed succeeded in reproducing the results of the direct analysis.

(vii) Number of components

The most difficult problem in any application of principal components is usually to decide how many components to use. For present purposes this can be defined as that which produces the minimum RMS residual of erroneous observations. The results of analyses using a variable number of stations and days showed that, to a first approximation, the best number of components could be expressed as a proportion of the number of stations used. This is because the complexity of the patterns being matched depends mainly on the number of stations used, as is evidenced by the figures presented in fig 5. The linear relationship will hold only so long as the best number of components is much less than the number of (independent) days since, if less than the number of stations, this places an upper limit on the number of components required to account for all the variance.

The more highly correlated the data, the fewer are the number of components that need be used. A consequence is that for minimum temperature the best number of components may be set at 6% of the number of stations, while for the less correlated sunshine data the figure rises to 9%. The number of sunshine stations used in this investigation, however (197), fell far short of the complete network (350).The shorter station separations for the full network provide more highly correlated data so that the best number of components may fall to 8% of the number of stations. The best number of components to use may therefore be estimated as 600 x 6% = 36 for temperature and
350 x 8% = 28 for sunshine.
The minimum in the relationship between the magnitude of the residual and the number of components is very flat, however. When only half the optimum number of components are used, the RMS residuals of erroneous observations increase by only 10%. This infers that the performance of the current operational system could be improved by 10% by doubling the number of components used from 15 to 30.

(viii) Internal consistency between elements

The analyses described so far have treated each element in isolation. Better results should be obtainable from analyses which take into account the consistency between elements. To do this properly would require a 3 dimensional analysis - stations x days x elements, whereas principal component analysis is essentially 2 dimensional. It is possible, however, to match the observations on a day against the components for several elements simultaneously. This reduces the deleterious effect of an error in just one element on the matching. This procedure was applied to maximum, minimum, dry bulb and wet bulb temperatures, a group of elements which are measured in the same units and which are available at the same network of stations.

The result of a single multi-element regression is that the weightings of each component on a given day are constrained to be the same for all 4 elements. As a consequence, the weightings of each component will be sub-optimal for any given element and a greater number of components will be required to achieve the same reduction in variance as that obtained from a single element regression. Against this, however, if there is an error at a station in just one element, then the correct values for the other 3 elements will reduce the fitting of the error at that site.

Another alternative is to replace the components from 4 single element analyses by those from a single 'multi-element' analysis. This can be achieved by using days as variables and increasing the number of cases from 600 to 2400 (say). The result is similar to that for 4 single element analyses except that the spatial patterns are modified slightly. The patterns for minimum temperature will no longer be optimal for minimum temperature alone, but the combination of all 4 patterns will be optimal when taken over all 4 elements.

A multi-element regression of data for 1978 was carried out on scores derived from a multi-element principal component analysis for 1973 and the results are presented in table 6. The RMS residuals of erroneous observations have been computed up to the 21st and 39th components for single and multi-element analyses respectively. When averaged over all 4 elements, the RMS residual of 1.02 deg C from 39 components of the multi-element analysis can be compared with 1.00 deg C from 21 components for a single element approach. It therefore appears that the reduced fitting of errors has not been able to overcome the disadvantage of the weighting of the components being constrained to be the same for all elements.

(viii) Seasonal variations

Seasonal variations in the RMS residuals of erroneous observations for minimum temperature using 15 components are displayed in fig 6a. The diagram also contains the standard deviation of the input data in the form of either station, daily, or local anomalies. It can be seen that the residuals are slightly smaller in summer than in winter, but that this is a weak reflection of the input data.

Seasonal variations in the residuals of daily sunshine based on 15 components are displayed in fig 6b and range from 0.8 hours in December to 1.7 hours in June, with an annual mean of 1.3 hours. The seasonal variation can be seen to be directly linked to the changing duration of maximum possible sunshine.

(ix) Geographical variations

Geographical variations in the RMS residuals of all observations for minimum temperature based on 15 components are displayed in fig 7. The local detail is unreliable because it depends on the site characteristics of individual stations, but in general the variations are as expected, with values ranging from 1.5 deg C in the data sparse Scottish Highlands to 0.7 deg C in parts of Northern Ireland and Southern England.

For daily sunshine in June, fig 8a reveals a similar dependence on the station network, with values ranging from less than 1.5 hours over much of England to over 2 hours over N.W. Scotland and parts of Wales, Western England, and the East Anglian coast. In December, however, fig 8b shows that geographical variations are much smaller, with values close to 0.8 hours. The reason for this is that in winter there are a large number of sunless days for which estimates are easy to make, and that the number of these days is greater in data sparse areas. On days with reasonable amounts of sun, a distribution of residuals similar to that in June, with highest values in the more data sparse areas, is expected.

7. Estimation by near neighbours

There are 4 main steps in the estimation of observations by a near neighbour scheme:-

(i) The selection of the neighbouring stations to be used.

(ii) The production of preliminary estimates from each of the neighbours.

(iii) The scanning of these estimates for errors.

(iv) The combination of the preliminary estimates to form a final value.

These are described below with particular reference to the procedures already used for rainfall.

The selection of neighbours

For rainfall the 8 nearest neighbours are selected subject to the proviso that no more than 2 may be drawn from any one quadrant. In the present work the 8 most highly correlated neighbours were chosen. If the correlation decay is isotropic, as is likely to be the case for sunshine and maximum temperature over simple terrain, the 8 most correlated neighbours will also be the nearest. Under these circumstances the stations are likely to be uniformly distributed and the correlation-based neighbours will be fairly evenly spread with respect to direction. For minimum temperature, and other elements in more complex terrain, a selection criteria based on correlation (r) is likely to prove superior to one based on distance. This superiority, however, is dependent upon the correlations being stable. If one years' data containing 120 independent values are used, for example, and the true correlation is 0.9, then the standard error of the correlation is 0.04. Hence more than one years data are required to obtain stable correlation coefficients.

The regression step

The conversion of observations from one station to another have conventionally been made with the aid of percent of annual average for rainfall, percent of monthly average for sunshine, and difference in monthly average for temperature. These simple conversion factors have been adhered to since more complicated relations were not expected to improve the estimates. A linear regression derived for a calendar month in the calibration period, for example, could not be expected to apply to the 'current' month. The conversion factors were treated in the following manner:-

(i) The averages used were either those for the period 1951-80 or were obtained from calibration periods of varying length.

(ii) The monthly ratios or differences were not used directly, but were meaned or smoothed over all months (using a 7 point binomial filter).

The scanning step

Erroneous observations at neighbouring stations will degrade the accuracy of the estimates derived from them and must be eleminated if possible. The following simple scheme is used to scan the estimates for likely errors.

The highest and lowest of the 8 estimates are excluded, the mean and standard deviation of the 6 central estimates are computed, and thresholds of ± 3 standard deviations from this mean value are evaluated. Each of the 8 estimates are then accepted or rejected according to whether they fall inside or outside of these thresholds.

The scheme can be expected to work well provided the thresholds are calculated from error-free data. This means that it can cope with at least one error per 8 neighbours and 2 if they are in the opposite sense. This should prove sufficient for the expected error frequencies of 1 in 100 for temperature and 1 in 190 for sunshine. The technique could be adapted to cope with an increased frequency of errors. If errors were common, for instance, more neighbours could be used and the mean and standard deviation calculated from only the central half of the estimates. Alternatively, a black list of poor stations could be maintained to prevent their acting as neighbours (this is done for rainfall). The non-normal distribution of estimates results in a poorer performance of this technique for sunshine than for temperature.

A threshold of 3 standard deviations (rather than 2, say) was chosen to counteract the tendency of standard deviations computed from a small number of central values to underestimate the true value. The rainfall quality control employs a threshold of 1.75 standard deviations, but this is because the highest and lowest of the estimates are included in the calculation of the standard deviation, the value of which is thereby inflated by errors.

The combination step

The accepted estimates from the individual neighbours are combined by attaching various weights to them. For rainfall the inverse of the square of the distance is used; here an inverse square law in correlation space, ie $1/(1-r)^2$ has been chosen. For minimum temperature, the typically produces a situation in which the 8th ranked neighbour is assigned a weight about one-fifth of that attached to the first neighbour.

Results

The RMS residuals of minimum temperature and sunshine obtained using the near neighbour scheme described above are presented in table 7. Since the quality of the estimates is unaffected by whether or not there is an error at the test station, there need be no distinction between correct and erroneous observations - their RMS residuals will be the same. Values are seen to lie around 1.0 deg C for minimum temperature and 1.1 hours for sunshine, and are therefore similar to those obtained using principal components. It made very little difference whether the differences or ratios were meaned over all months or whether the monthly values were smoothed to retain a seasonal variation. Results based on 30 year averages were matched by the use of a single year calibration period, and small improvements were obtained when the length of the calibration period was extended to 4 years.

Seasonal variations were very similar to those obtained for principal components, but the geographical variations, while similar, are more pronounced, and are displayed in figs 9 and 10. RMS residuals vary from 0.6 to 2.0 deg C for minimum temperature and 1 to 3 hours for June sunshine, much larger ranges than those obtained from principal components. Principal components therefore have the advantage of performing better at the more isolated stations.

Comparison with principal components

There is now sufficient evidence to indicate that the performance of the principal component technique gains relative to the near neighbour approach as the inter-station correlation decreases. It is demonstrated by the fact that where the station network is good then near neighbours hold the advantage but that where it is poor principal components are superior. The failure of principal components to improve on the near neighbour estimates for the relatively poorly correlated sunshine data is attributed to the non-normal distribution of daily sunshine. Hence it is postulated that principal components gains over near neighbours as the data become more poorly correlated but loses as they become less normally distributed.

8. Estimation from the current month

A requirement of an operational quality control system which has to be recognised is the need to cope with new stations. For these stations past data is unavailable for calculation of principal components or near neighbour regressions, and substitute information has to be derived from data for the 'current' month, ie the month which is being quality controlled. For the principal component technique this is almost entirely disadvantageous since the components have to be estimated, leading to a degraded performance. For the near neighbour technique, however, the approach has one big disadvantage, namely that the data to be estimated are not independent of those used to develop the regressions.On this account, the 'current month' approach is worth considering as a general alternative to the use of past data. A disadvantage, however, is that the observations available for forming the regressions contain errors so that the data have to be scanned and the likely errors removed. It follows that there are 2 classes of observation - the majority which have been used to develop the regressions and are easily estimated, and a minority, which include the likely errors, which have been excluded from the regressions and which are difficult to estimate. In compiling RMS statistics of the residuals, it therefore becomes necessary once again to distinguish between those for correct and erroneous observations.

8.1 Near Neighbours

The methodology of the current month technique is essentially the same as that used with past data but an extra first step is provided to remove or 'trim' likely errors from the data. The selection of neighbours and regression steps are also modified.

The trimming step

As the aim is to provide observations suitable for the calculation of regressions and correlations between stations, the data were treated in station pairs. The differences between the observations at the stations were calculated, and the values which were trimmed or excluded were those with the t highest and lowest values of the differences. A criterion based on station difference was considered suitable for daily sunshine as well as temperature. Station pairs with 6 or more missing differences were not considered for further analysis. The value of t used was set at 3 (leading to the rejection of about 20% of the data in a month), but for temperature, for which the data analysed in this report consisted of only 10 values per month (ie one observation from every third day), the value was set at 1. The number of observations to be trimmed should be related to the expected frequency of errors, so a black list of poor quality stations with increased values of t could be used.

A variant of the above scheme was tried in which the mean and standard deviation of the 'trimmed' differences were calculated. Thresholds of ± so many SD's from the trimmed mean were calculated and each pair of observations were accepted or rejected according to whether their differences fell inside or outside the prescribed tolerance zone. This routine failed to improve on the procedure outlined above, however, and so the simpler scheme was retained.

Selection of neighbours

The calculation of a full correlation matrix, its sorting, and the subsequent selection of neighbours is an expensive operation. If the neighbours are selected from data in a 'calibration period' this task need be performed no more than once per year and the results written to disc for reading by the operational monthly ppograms. This procedure is analogous to the reading of pre-calculated scores or loadings in a principal component scheme. The need to include a full selection of neighbours scheme for all stations in the operational monthly program would render the current month technique for all stations prohibitively expensive. A simple compromise is to provide the program with a pre-selected set of 16 neighbours (the nearest, for example) from which the best 8 in the current month may be chosen. This is a relatively inexpensive task.

The regression step

The constant difference and constant ratio conversion factors used in the past data approach were retained. Also tested, however, were conventional linear regression relationships made suitable by the fact that they would be operating only on dependent data. The form of regression in which the slope is set equal to the ratio of the standard deviations was found to be better than that in which the slope equals the product of this ratio with the correlation.

For daily sunshine the constant ratio is not an ideal conversion factor because a completely sunny day at one station will (unless the ratio equals unity) predict a value which will exceed, or never reach, the maximum possible amount at the other station. Although one station may be sunnier than another, the relationship between the stations will be fixed at zero and the maximum possible daily sunshine. Various attempts were made to devise algorithms which satisfied these constraints, but they all failed to improve on the estimates provided by the constant ratio conversion factor.

Results

There are several gradations between the 'past data' and 'current month' techniques. The selection and weighting of neighbours can both be made to depend upon correlations determined from either the calibration period or current month. The RMS residuals associated with a variety of regression techniques and neighbour and weighting routines are presented in table 8. Only the option based entirely on the current month is available for new stations. The RMS residuals taken over all and erroneous observations diverge steadily as the options involved become more completely centred on the current month. For minimum temperature the RMS residual of 1.0 deg C then rises to 1.17 deg C for erroneous observations but falls to 0.71 deg C for the remainder. From these figures it can be deduced that the current month technique is providing a similar standard of quality control to that obtained from the use of past data.

8.2 Principal Components

In the current operational quality control, the missing component scores for new stations are estimated from data for the current month by program BMDPAM of the BMDP suite of statistical software. In this section various options in the BMDPAM program are tested and the accuracy of the residuals for minimum temperature assessed. The estimated scores are biassed in favour of the data in the current month and so, as for the near neighbour technique, better estimates of the majority of observations can be made than if the component scores were based entirely on independent data. The estimated component scores are also biassed towards the error, however, making their detection more difficult.

The current implementation of BMDPAM is as follows:-

(i) Cases are assigned to stations, and variables to the 31 days in a month and the 15 component scores.

(ii) Multiple linear regression is based on covariances calculated from all available data (the ALLVALUE) option. Alternative covariances can be calculated using only complete cases or by using a maximum likelihood technique (options COMPLETE and ML respectively).

(iii) The off-diagonal covariances are multiplied by 0.9 by making use of a RIDGE parameter which simulates ridge regression. The effect of this is to reduce the standard errors of the regression coefficients at the expense of also reducing the value of the regression coefficients themselves.

Program BMDPAM was tested by witholding component scores from 15% of stations, this proportion corresponding to the number of new stations which might be expected to accrue over a period of 5 years. The provision of data for a given month was complicated by the fact that the temperature data extracted for this investigation represented every third day, giving only 10 days per month. A compromise was made by providing 20 days data taken from 2 consecutive months. This data was then contaminated by errors and together with the incomplete component scores submitted to BMJPAM which was then run with a variety of RIDGE parameters and covariance options.

Without ridge regression (ie RIDGE=1) all 3 methods of calculating the covariances gave similar results and those for the ALLVALUE option are displayed in table 9. The over-fitting of both the component scores and the weights attached to them leads to the smallest RMS residuals of erroneous observations of about 1.5 deg C being associated with just 3 components. If 15 components are used, the residuals have risen to 2.0 deg C, nearly twice that associated with the use of genuine scores.

When a RIDGE parameter is applied, the results are unaffected when the covariances are calculated using option ML, while those for ALLVALUE are similar to COMPLETE and are also displayed in table 9. The effect of the RIDGE parameter is to reduce the regression coefficients, and hence the component scores, towards zero. Since the component scores computed without ridge regression are based on the correct observations, this loss of variance leads to an increase in the residuals from the correct observations. For erroneous observations, however, the un-modified scores fit the errors rather than the true values and a modest loss of variance is capable of improving the estimate. A further convenience of ridge regression is that as the estimated scores approach zero they fail to make a contribution to the estimates of the observations which therefore become independent of the number of components used. The most appropriate choice of RIDGE is perhaps 0.5, for which residuals of correct observations are reasonably small while those of erroneous observations are close to their minimum and almost independent of the number of components used. For 15 components, this gives residuals of 1.13 and 1.44 deg C for correct and erroreous observations respectively, an improvement over the 0.93 and 1.78 deg C obtained with the current operational system (for which RIDGE = 0.9). These results, however, are still inferior to those obtained using a near neighbour scheme for which the corresponding RMS residuals are 0.71 and 1.17 deg C respectively.

9. Flagging routines

Introduction

The overall performance of any quality control procedure depends on

(i) the accuracy with which true values can be estimated on occasions of erroneous observations.

(ii) the number of errors which are not queried (type I error)

(iii) The number of invalid queries raised (type II error).

So far attention has been concentrated on item (i) and the RMS residuals of erroneous observations. The overall performance of a quality control system also depends on the RMS residual of correct observations, however, since this has a bearing on the number of invalid queries raised. As far as the raising of flags and the tuning of a technique is concerned, the residuals of erroneous observations affect the number of errors which are not queried, while the residuals of correct observations affect the number of invalid queries raised.

The number of errors which remain undetected is not an optimum parameter for assessing the merit of a quality control system because it depends on the distribution of errors. If there are a large number of small errors, for instance, it may not matter if they are not detected so long as the large errors are. Thus the proportion of error variance detected would be a better measure of the effectiveness of a quality control system. No absolute measure of merit is possible, however, since subjective judgement is required to decide whether to aim for economy and minimise the number of invalid queries raised, or to go for quality and maximise the proportion of error variance detected. In this report therefore, the tuning of a technique is expressed by the simple expedient of quoting the numbers of valid and invalid queries raised as a fraction of the number of errors present.

Flagging procedures

A query will be raised whenever an observation differs widely from its estimate, and the best way of defining a large residual has to be found. On some occasions minimum temperatures (say) at neighbouring stations will be similar to one another and hence easily estimated, with a correspondingly low variance in the residuals. On other occasions, eg radiation nights, the differences between stations will be much larger with a corresponding increase in the range of the residuals. These differences can be taken into account by raising flags whenever a residual exceeds a specified number (S) of standard deviations of the residuals. It is also advantageous to supply an absolute threshold (D) to prevent small differences from being queried. Thus flags are raised whenever a residual exceeds the larger of D or S. The current quality control uses only S, but it is set at such a high value (3.25) that small values of D will have little impact.

A number of options are available for the calculation of the mean and standard deviation (SD) of the residuals:-

(i) The mean and SD of the residuals could be calculated from all stations in the UK. Alternatively, the country could be divided into 10 districts and the mean and SD calculated for each area separately. High and low variances of residuals in different districts could have adverse affects on the efficiency of the flagging in each district. This was the philosophy adopted for the current quality control. Taking the argument further, only the residuals of stations in small groups of counties could be used to decide which observations should be queried. This makes the flagging procedure similar to that in a typical near neighbour scheme.

(ii) When the country is divided into smaller areas, the mean of the residuals in each region will no longer be zero. There are likely to be systematic errors in the prediction of the principal components, and these can be removed by using the mean and SD of the sample of residuals being used. This is the course taken by the current quality control procedures. Bearing in mind that some of the data contain errors, however, this sample mean and SD may be misleading, and so the option of setting the mean of the residuals to zero, and calculating the SD about zero, has been included.

(iii) The effects of errors can be further reduced by scanning the residuals. They can be ranked, the highest and lowest 10% of values excluded, and the mean and SD calculated from the remaining values.

All these options were tested on the principal component scheme. For the near neighbour technique, only the residuals from the nominated neighbours were used in the flagging routine, but there is no reason why the procedures described above could not be used. It should be noted that dividing the country into districts and then raising queries according to the mean and SD of the residuals in those districts, is a way in which the 'flagging routines' can effect the overall accuracy of the technique. The RMS residuals presented so far were calculated about a mean of zero; if systematic errors over a district can be removed, then this increases the accuracy of the technique.

The 'best' number of components to use has so far been taken to be that associated with the smallest residuals of erroneous observations. The systematic errors in residuals will gradually decrease with the number of components used, and so their removal will have the biggest impact for a small number of components. This will therefore lead to a decrease in the best number of components to use. The residuals of correct observations also affects the performance of a quality control system, however, and their magnitude decreases steadily as the number of components increases. This effect will therefore lead to an increase in the best number of components to use.

Results for minimum temperature

Principal components based on daily anomalies and a T mode analysis were computed for the whole UK for 1973 and used to estimate minimum temperature in 1978. Queries were raised according to the following options:-

(i) the mean and SD of the residuals were calculated from stations in districts about 1/10th the size of the UK, and in smaller groups of counties.

(ii) the standard deviation of the residuals was calculated about an assumed mean of zero, or else both mean and SD were calculated from the sample (ie the systematic error was removed).

(iii) the mean and SD of the residuals were calculated with and without scanning (trimming) for the residuals.

The numbers of undetected errors and invalid queries were standardised by expressing them as a percentage of the number of errors and a single assessment of performance made using a scoring system in which an undetected error was rated as 10 times as important as an invalid query

ie performance score = errors missed +0.1 x Invalid queries

Thus if the no of invalid queries was equal to the number of errors (ie 100% of the no of erors) and 20% of these errors were undetected then

performance score = 20 + 0.1 x 100 = 30.

Results are presented for the flagging thresholds (D and S) which gave the lowest performance scores.

Too much importance should not be attached to this arbitrary scoring system. The performance of the options change little whatever the relative importance of type I to type II errors, but a single measure of assessment is useful for presenting the results, which are displayed in table 10.

In table 10, consider first residuals analysed by district, and in particular the option in which the residuals are not trimmed but their mean is assumed to be zero. These choices correspond to those made in previous section in which the smallest residuals of erroneous observations were found to occur when the number of components equalled 22 (6% of the 373 station used). As noted earlier, the RMS residual for correct observations continues to decrease with increasing components, causing the best performance as assessed here to be reached after 30 components. Removing the systematic error in the residuals, however, produces a marked improvement in the results for a small number of components, making the performance insensitive to the number of components used. Trimming the residuals, or analysing them by county rather than district, made little difference.

The tuning of the quality control system by variation of D and S is illustrated in fig 11 for both the principal component and near neighbour techniques (past data and current month options). The principal component results are based on the options used in the current operational system, ie 15 components in which the residuals are not trimmed but their systematic error is removed. It can be seen that the near neighbour technique, and in particular the current month option, holds slight advantages over the principal component approach, but in general, about 80% of errors can be detected by raising twice as many queries as there are errors. A point of interest is the extent to which the flagging threshold of S = 3.25 used in the current system is geared to economy. The number of invalid queries raised is very small, but this is only achieved at the expense of 40% of the errors remaining undetected.

Results for sun

The performance of a quality control procedure, when assessed in terms of error detection, improves as the frequency of errors approaches 50%. This phenomenon is discussed in the next section but it does mean that in order to make comparisons of the performance of the quality control of sun relative to that of minimum temperature it is necessary to assume the same frequency and distribution of errors. The distribution of sunshine errors in not dissimilar to that for temperature and has been left unchanged, but the frequency of errors in this section is assumed to be 1 in 100, ie the same as that for temperature. The frequency of sunshine errors found by the quality control staff and ascribed to areal quality control was observed to be only 1 in 190. It was also found, however, that many sunshine errors were not being detected so it may well be that the true rate of error for sunshine is the same as that for temperature.

The performance scores obtained from an S mode principal components analysis based on daily anomalies for 1979 and tested on data for 1983, are displayed in table 11. The main finding is that analysing residuals by a small group of counties rather than by district is harmful, a fact ascribed to the non-normal distribution of sunshine. This renders the mean and SD of residuals calculated from only a small number of stations (as low as 6) a poor guide to the identification of errors. The non-gaussian distribution of residuals increases the sampling variability of the mean and SD, and this will decrease as the number of stations considered increases.

Other points of interest to note from table 11 are as follows:-

(i) Earlier in the report, when residuals were analysed by district, untrimmed, and with the systematic error not removed, it was found that the smallest residuals for erroneous observations were associated with 18 components. For those same options, (8% x 197 stations) the contribution of the residuals for the correct observations, causes the performance measured in terms of error detection to improve up to the largest number of components (25) examined.

(ii) Removing the systematic error yields only a small improvement for a small number of comonents when the residuals are analysed by district.

(iii) Trimming the residuals leads to small improvements in most cases.

The tuning of the quality control of sunshine by varying D and S is illustrated in fig 12. First consider the points denoted by circles, which represent results from a near neighbour technique with an error frequency of 1 in 190. These represent a very poor standard of quality control, with over 6 times as many queries as errors needed to detect 70% of the errors. The points denoted by squares represent a near neighbour technique in which the frequency of errors is increased to 1 in 100. A dramatic improvement (to be discussed later) is indicated because although the number of queries raised is slightly more than before, the number of invalid queries is practically halved when expressed in terms of the number of errors. The points denoted by triangles represent results obtained using 15 components and with residuals analysed by county, untrimmed, and with the systematic errors removed. A modest improvement, to be discussed later, is evident. The points denoted by diamonds are for a PCA in which the options are the same as before except that the residuals are analysed by district rather than county. This combination of options is not that which gives the best performance but that which corresponds to those used by the current quality control. The improvement of the analysis by district rather than county, discussed earlier, is evident, and the number of queries required to detect 70% of the errors is now less than two and a half time the number of errors.

The reason why principal components with residuals analysed by counties performs better than a near neighbour technique in terms of error detection is thought to be as follows. Although the residuals for principal components and near neighbour techniques have a similar RMS value, those for principal components have less scatter, the residuals being smaller for more isolated stations. It is these isolated stations which generate the most queries and so an improvement in the estimates for these stations helps to reduce the number of invalid queries raised. These arguments apply to temperature as well as sunshine, but the different rates of correlation decay cause the effect to be much greater for sunshine than temperature.

10. Effect of the frequency of errors on their detection

A decrease in the frequency of errors is associated with an improvement in the accuracy with which estimates of erroneous observations can be made, but this is only slight as is illustrated in table 12. Of more importance is the fact that as the error frequency decreases, the detection of those errors becomes progressively more difficult, and the reasons for this are described below.

Consider a situation in which a set of observations contain some errors of ±4 deg C and that the accuracy (RMS residual) with which the true values can be estimated is 1 deg C for both correct and erroneous observations. If as many as 50% of the observations are in error then the distribution of residuals is illustrated in fig 13a (where these appear to be twice as many correct as erroneous observations because the negative residuals are not displayed). There is good discrimination between the correct and erroneous observations, with very little overlap between the two distributions. If S is set at 2.5, for instance, then 7% of the errors would remain undetected while the number of invalid queries raised would be minimal.

Suppose next that the frequency of errors was decreased to 1 in 10, a situation illustrated in fig 13b. The relative increase in the number of correct observations has increased the overlap between the residuals of the correct and erroneous observations. By setting S to 2.5 the proportion of undetected errors remains at 7% but the number of invalid queries has increased to 12% of the number of errors. If the frequency of errors is decreased to 1 in 100 (fig 13c), the overlap between the two sets of residuals is increased still further - for s set to 2.5, the number of invalid queries has increased to 1.24 times the number of errors.

It should be borne in mind that as the frequency of errors increases, the more difficult it is to make estimates and RMS value of the residuals increases. Furthermore, a decreasing frequency of errors is associated with a decrease in the total number of queries raised; it is only when expressed as a fraction of the number of errors that a drastic increase in the number of invalid queries occurs. Nevertheless, fig 13 does demonstrate that, all other things being equal, it becomes progressively more difficult to discriminate between correct and erroneous observations as the proportion of errors decreases below 50%.

The situation illustrated in fig 13c approximates to that pertaining to the quality control of minimum temperature, although in the latter case the errors are not fixed at 4 deg C and so the spread of the residuals of erroneous observations will be greater than that indicated. Nevertheless, it can be seen how the current threshold of S = 3.25 is close to the point at which an observation is just as likely to be in error as it is to be correct. If S is decreased below this value the proportion of errors detected increases but the number of invalid queries rises rapidly.

The threshold to be chosen for raising flags depends on a management decision as to whether to aim for economy or quality. For England and Wales areal quality control occupies 1/6th of the time of 4 persons involved in all aspects of the quality control of temperature and sunshine, and most of this is involved in the oral or written checking of observations. There is therefore not much extra staff effort involved in aiming for quality rather than economy. For temperature it seems reasonable to generate twice as many queries as errors, corresponding to the setting of D = 2.5 and S = 2.0 (as opposed to S = 3.25). Most of the extra errors detected will be small, however, mainly between 2.5 and 3.25 deg C. So far it has been tacitly assumed that once a query has been raised, the human scrutineer is able to make the correct decision as to whether an observation is in error or not. With the small errors under discussion the success rate may be small and will depend on the experience of the scrutineer. As junior staff are commonly employed in this role, there may be little point in reducing S below 3.25.

11. Application of principal components to quality control of rainfall

The question arises as to how well a principal component scheme might fare when applied to rainfall, and whether it could improve on the current near neighbour technique. The following factors are relevant.

(i) Although the correlation decay for rainfall is much more rapid than for sunshine or temperature, the density of the gauge network is such that the highest inter-station correlations are generally greater than 0.9 and sufficient for a near neighbour scheme to operate reasonably well (O'Connell et al, 1977). A notable feature of the correlation decay for rainfall is its pronounced seasonal variation, with values being several times greater in summer than in winter (Stol, 1972). A consequence is that the near neighbour technique is expected to perform well in densely gauged areas and in winter, but to offer plenty of scope for improvement in data sparse areas and in summer.

(ii) Daily rainfalls are highly skewed and, as has been seen with sunshine, this is likely to pose severe problems for the principal component technique. Unlike sunshine, the skewness associated with rainfall will always be in the same sense, and hence can be considerably reduced by the application of a simple square root transformation (say). Such a transformation will be unable to deal with that part of the skewness caused by a large number of zeros, however, and the frequent estimation of negative rainfalls seems inevitable. The ability of principal components to improve on a near neighbour scheme for rainfall therefore seem likely to be severely compromised by its inability to cope with skewed distributions.

12. Conclusions

The principal component and near neighbour techniques provide similar standards of quality control with RMS errors of estimates of about 1 deg C for minimum temperature and 1.2 hours for daily sunshine. For error frequencies of 1 in 100 the number of queries required to detect 80% of the errors is 2 and 3 times the number of errors for minimum temperature and sunshine respectively. Principal components gains over near neighbours as the data become less well correlated and hence has the advantage of performing better at the more isolated stations. The near neighbour technique, however, produces the better estimates for new stations for which past data are unavailable. Non-normally distributed data are responsible for degrading both the standard of estimates obtained from principal components and the effectiveness of flagging routines based on only a small number of stations, as in a typical near neighbour sceme.

The current quality control system is close to being optimal but the following points are worthy of note:-

(i) The residuals of erroneous observations can be reduced by about 10% by doubling the number of components used from 15 to 30. However, the removal of systematic errors on the one hand, and the effect of the residuals of correct observations on the other, combine to produce a standard of quality control which changes little over a wide number of of components.

(ii) The system of calculating principal components from every third day over a 5 or 8 year period could be replaced by one based on using every days observations from only one year of data. The decrease in accuracy for established stations (about 5%) would be offset by a simpler updateable system in which the number of new stations were kept to a minimum (3%).

(iii) In the estimation of components for new station using BMDPAM, changing the ridge parameter from 0.9 to 0.5 should reduce residuals of minimum temperature from 1.8 to 1.5 deg C.

(iv) The current level of tuning of S = 3.25 is geared to economy by raising queries at a threshold at which observations are just as likely to be in error as they are to be correct. By generating twice as many queries as errors, the threshold could be reduced from 3.25 to 2.5 deg C, but the potential benefits will probably be lost through the difficulties experienced by the human scrutineer in correctly identifying these small errors.

The overall standard of quality control attained by use of the conventional station network may be considered as adequate for temperature but unsatisfactory for sunshine. This is a natural consequence of the reduced network of stations and greater correlation decay for sunshine compared with temperature. Further reductions in the network of sunshine stations are planned as a less than commensurate increase in the number of radiation stations takes place. It is therefore suggested that satellite data be used as the basis of an operational system for the quality control of a combined network of sunshine and radiation stations.

References

Bryant, G.W.	1979	Archiving and quality control of climtological data. Met Mag, 108, 309-315.
Hopkins, J.S.	1977	The spatial variability of temperature and sunshine over uniform terrain. Met Mag, 106, 278.
Kendall, Sir Maurice	1980	Multivariate Analysis. Griffin, London.

O'Connell, P.E., Beran, M.A., Gurney, R.J., Jones, D.A. & Moore, R.J. 1977 Methods for evaluating the UK raingauge network. Institute of hydrology Report No. 40, NERC, Wallingford.

Richman, M.B. 1986 To be published in J. Climatology.

Spackman, E.A. 1980 Areal quality control of daily climatological data using station factor scores. Met O 3 Tech Note No 6.

Spackman, E.A. & Singleton, F. 1982 Recent developments in the quality control of climatological data. Met Mag, 111, 301-311.

Stol, Ph.Th. 1972 The relative efficiency of the density of rain-gauge network. J. Hydrology, 15, 193-208.

FIG1 - DISTRIBUTION OF STATIONS FOR (a) TEMPERATURE AND (b) SUNSHINE.

Dotted lines denote the boundaries of 10 regions into which the U.K. was divided.

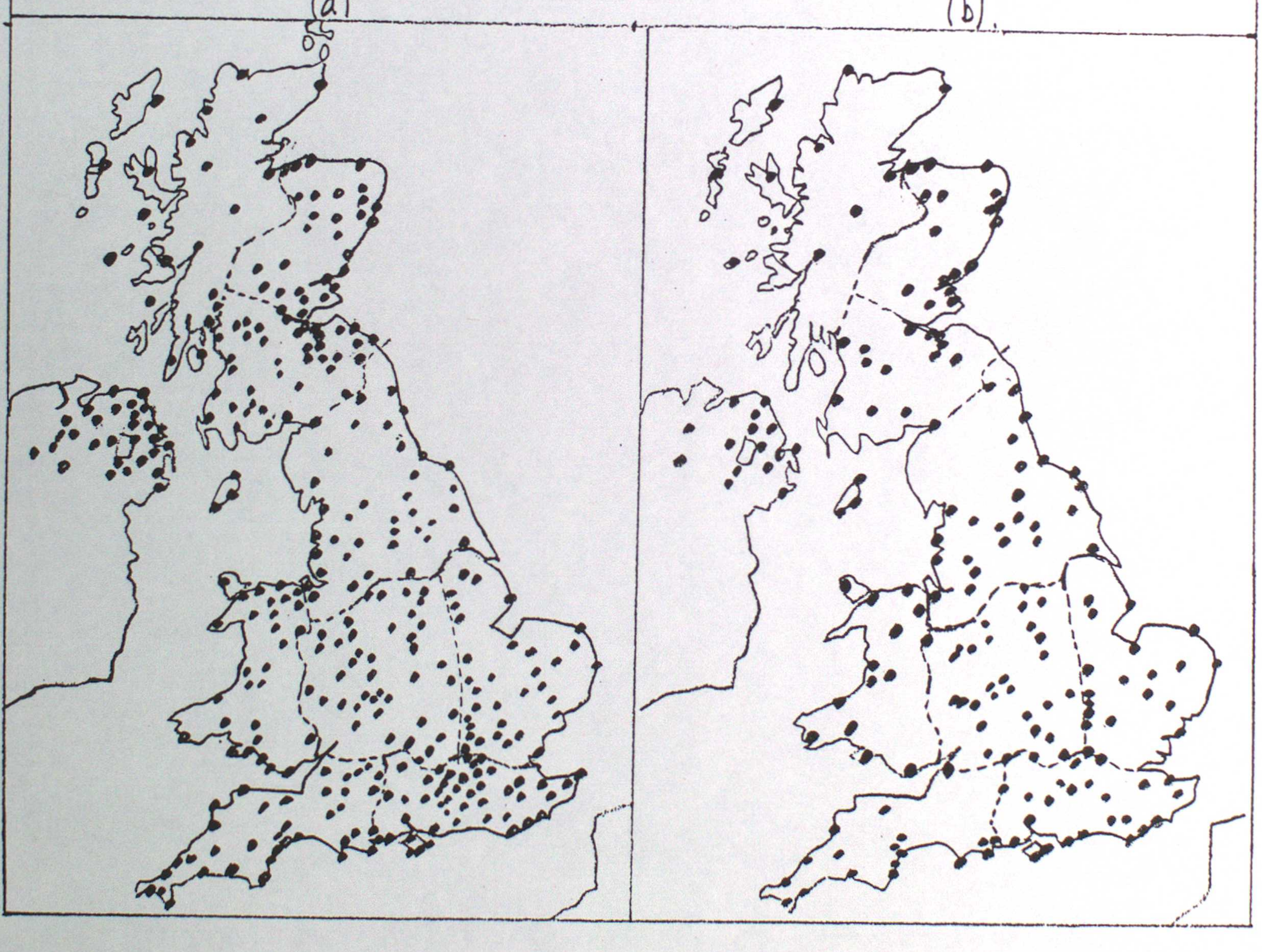


FIG 2 - HIGHEST INTER-STATION CORRELATIONS OF DAILY VALUES OF (a) MINIMUM TEMPERATURE IN 1973 AND (b) SUNSHINE IN 1979

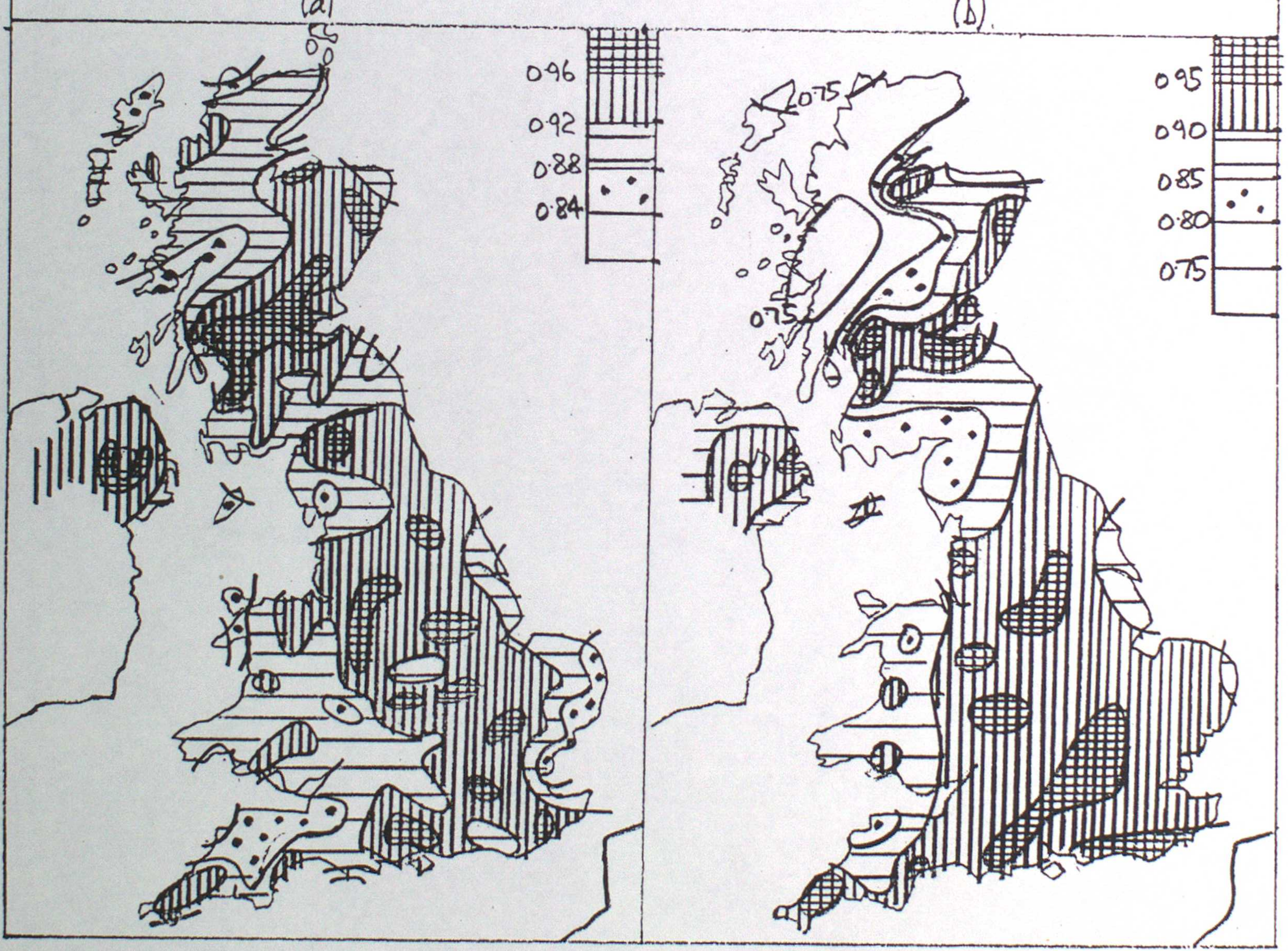
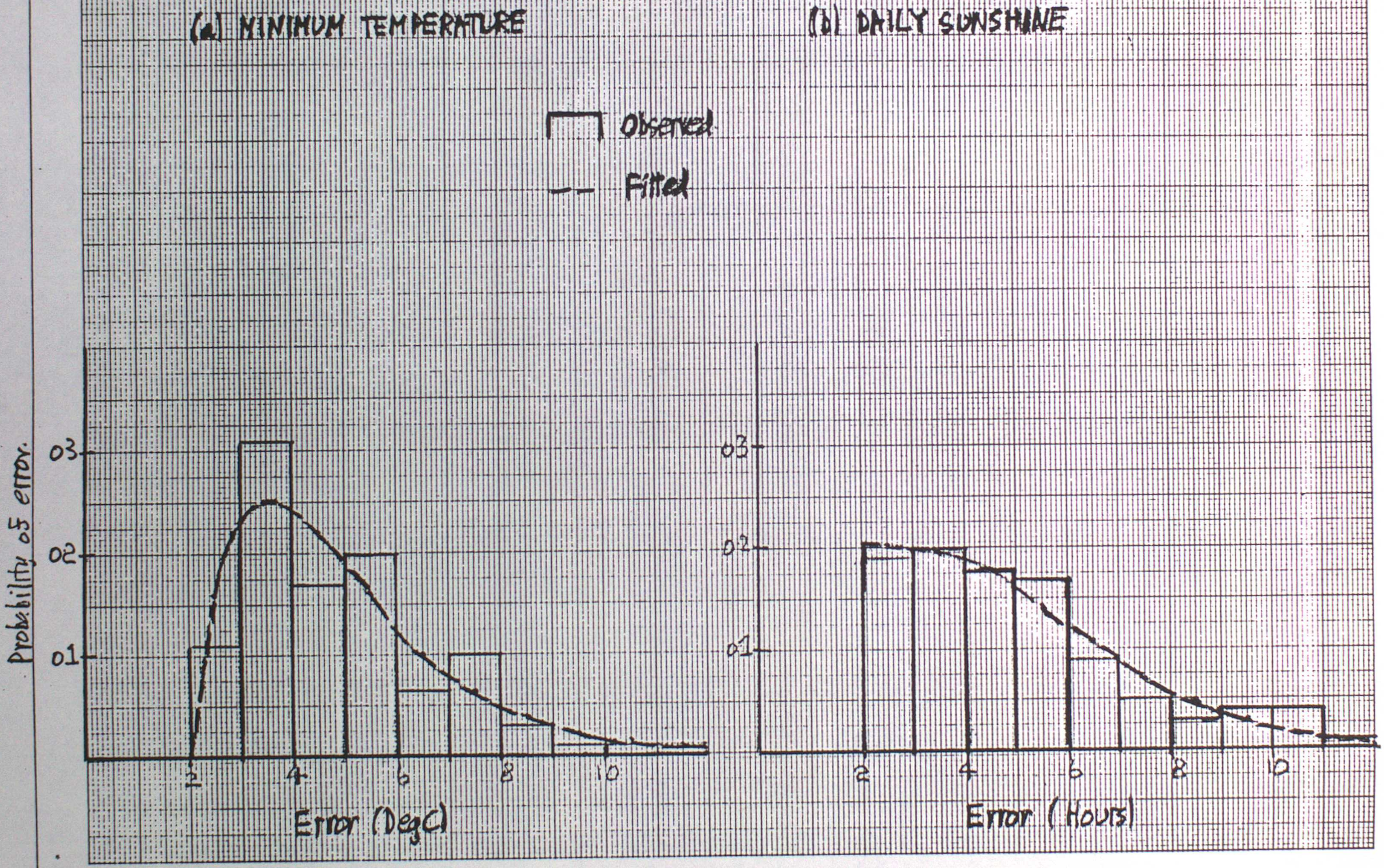


FIG3 - DISTRIBUTION OF ERRORS IDENTIFIED BY SPATIAL CONSISTENCY CHECKS



(a) MINIMUM TEMPERATURE

(b) DAILY SUNSHINE

Fig 4 – Use of a second PCA to form 'averages' of components obtained from previous analyses.

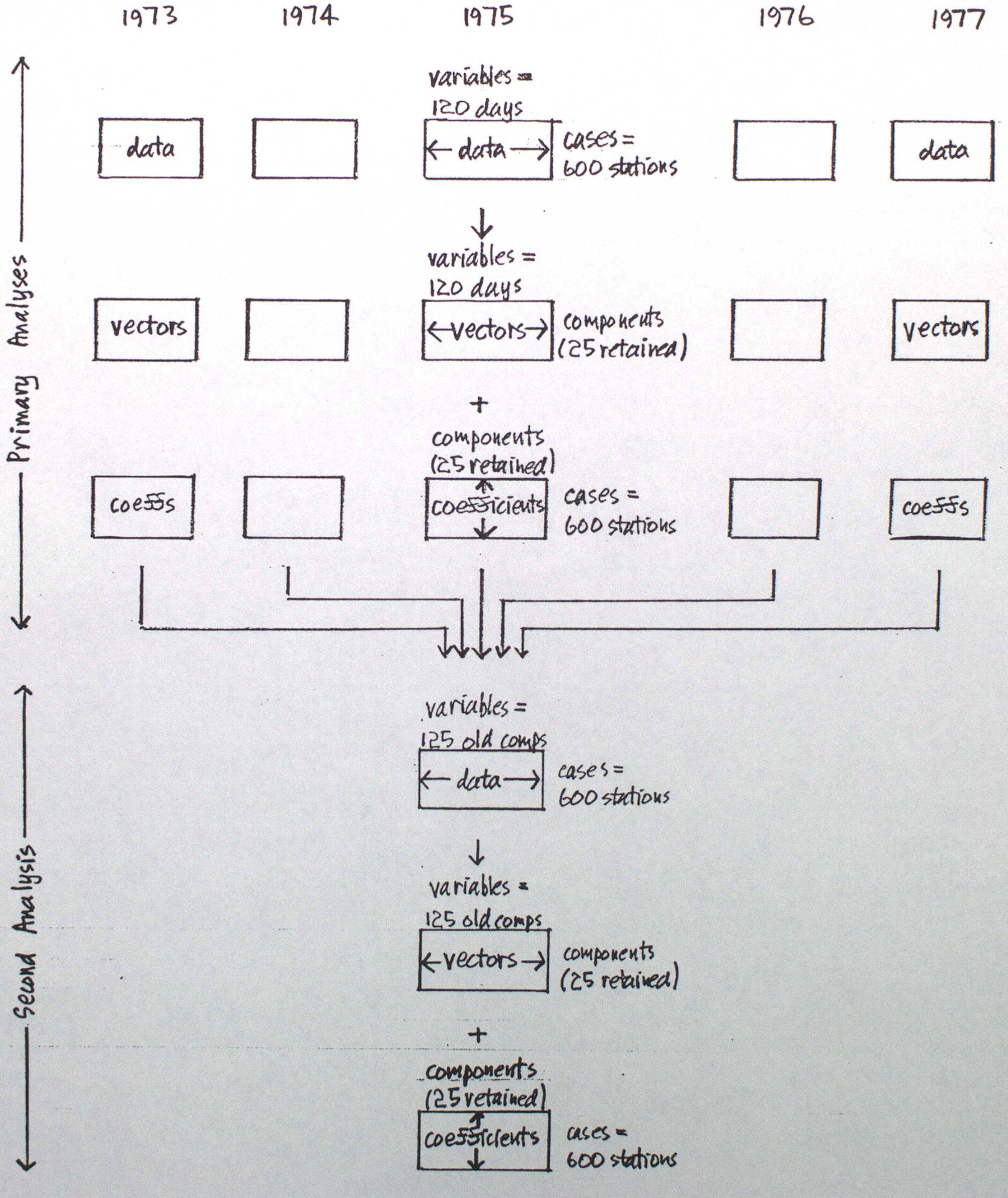


FIG 5 - DEPENDENCE OF THE BEST NUMBER OF COMPONENTS ON THE NUMBER OF STATIONS.

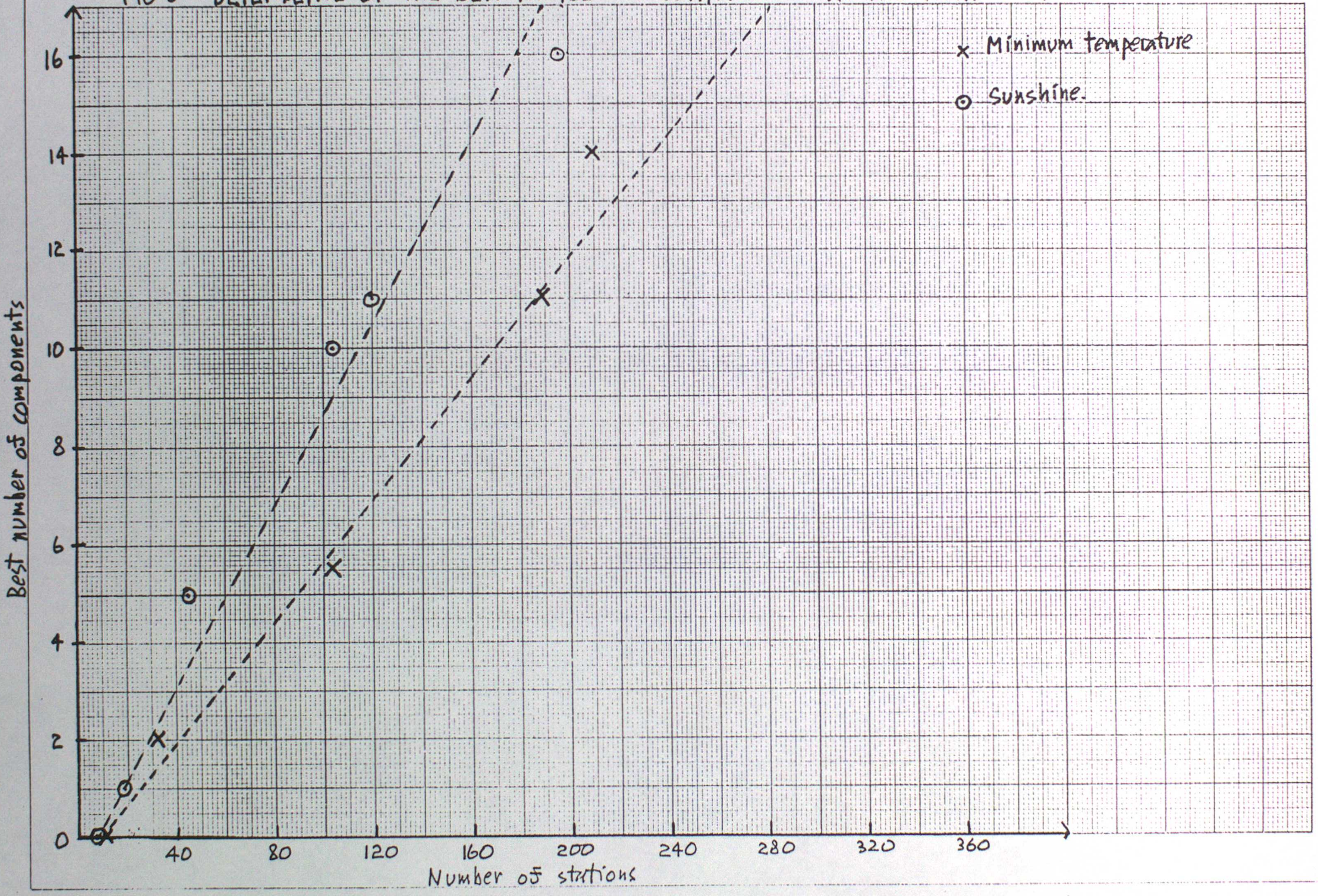
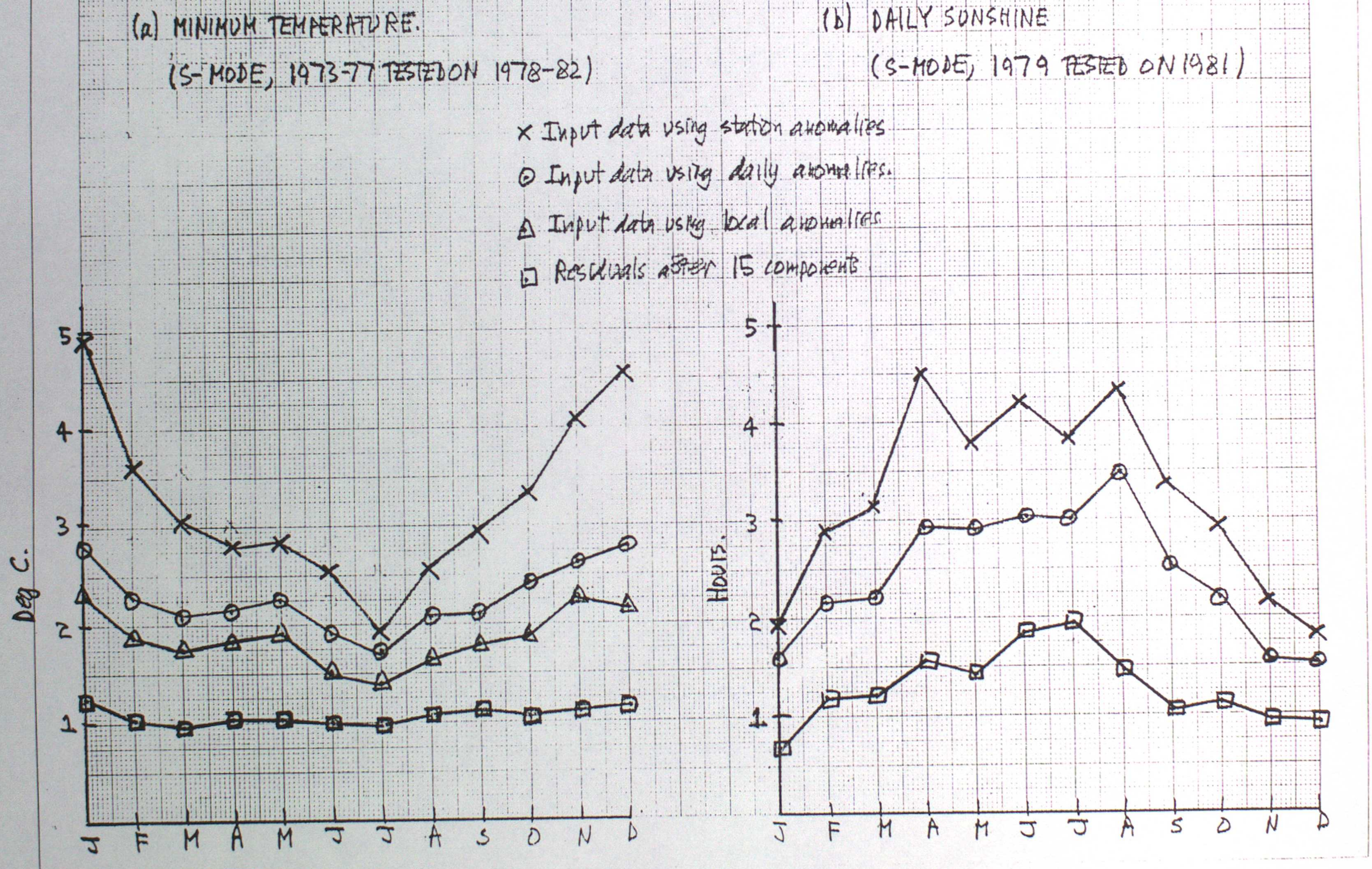


FIG 6 - SEASONAL VARIATIONS IN RMS RESIDUALS OBTAINED USING PRINCIPAL COMPONENTS.



(a) MINIMUM TEMPERATURE.

(S-MODE, 1973-77 TESTEDON 1978-82)

(b) DAILY SUNSHINE

(S-MODE, 1979 TESTED ON 1981)

FIG 7 - GEOGRAPHICAL VARIATIONS IN RMS RESIDUALS OF MINIMUM TEMPERATURES (DEG C) OBTAINED FROM PRINCIPAL COMPONENT ANALYSIS, (S-MODE: LOCAL ANOMALIES, 15 COMPONENTS, 1973 TESTED ON 1978)

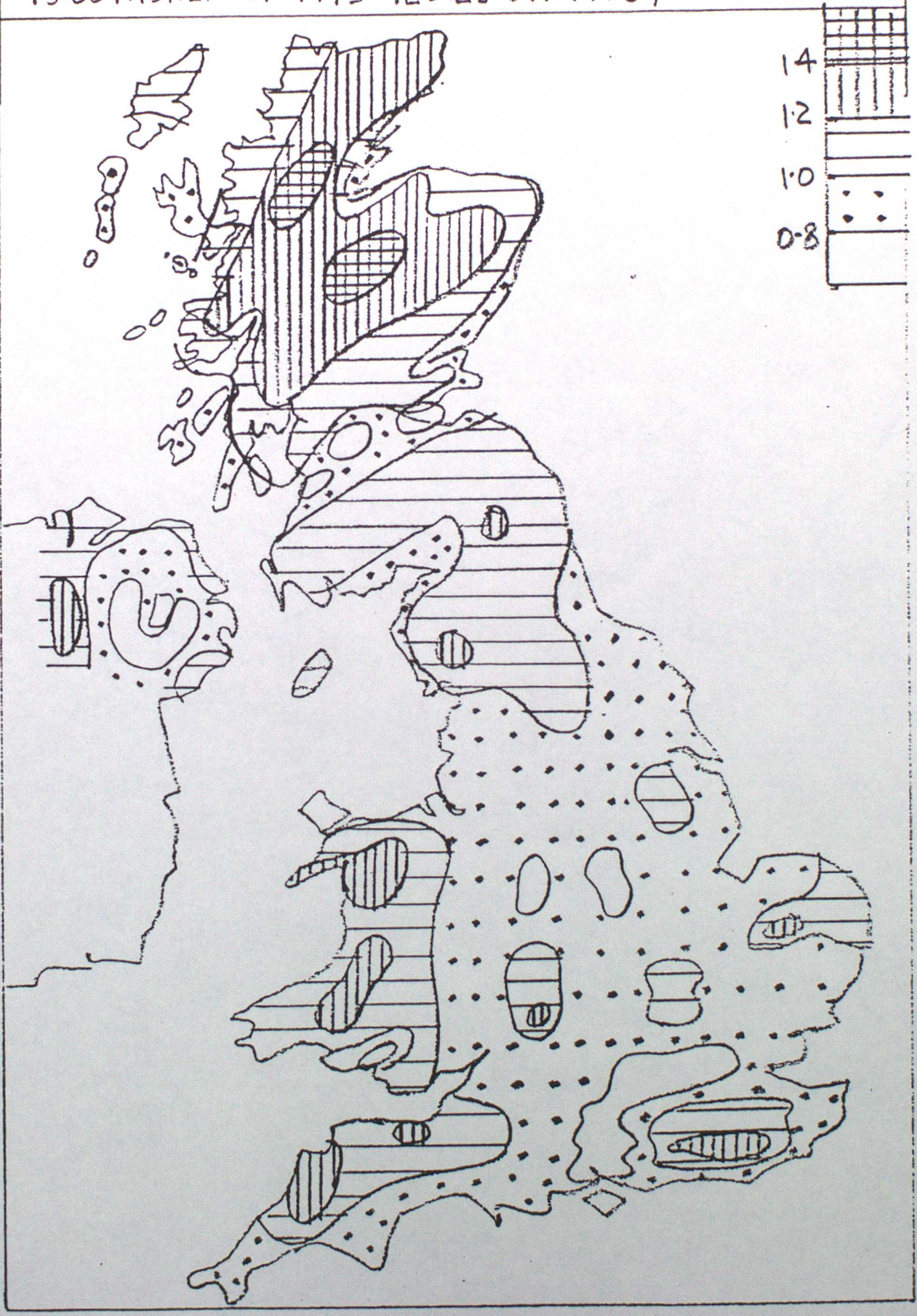
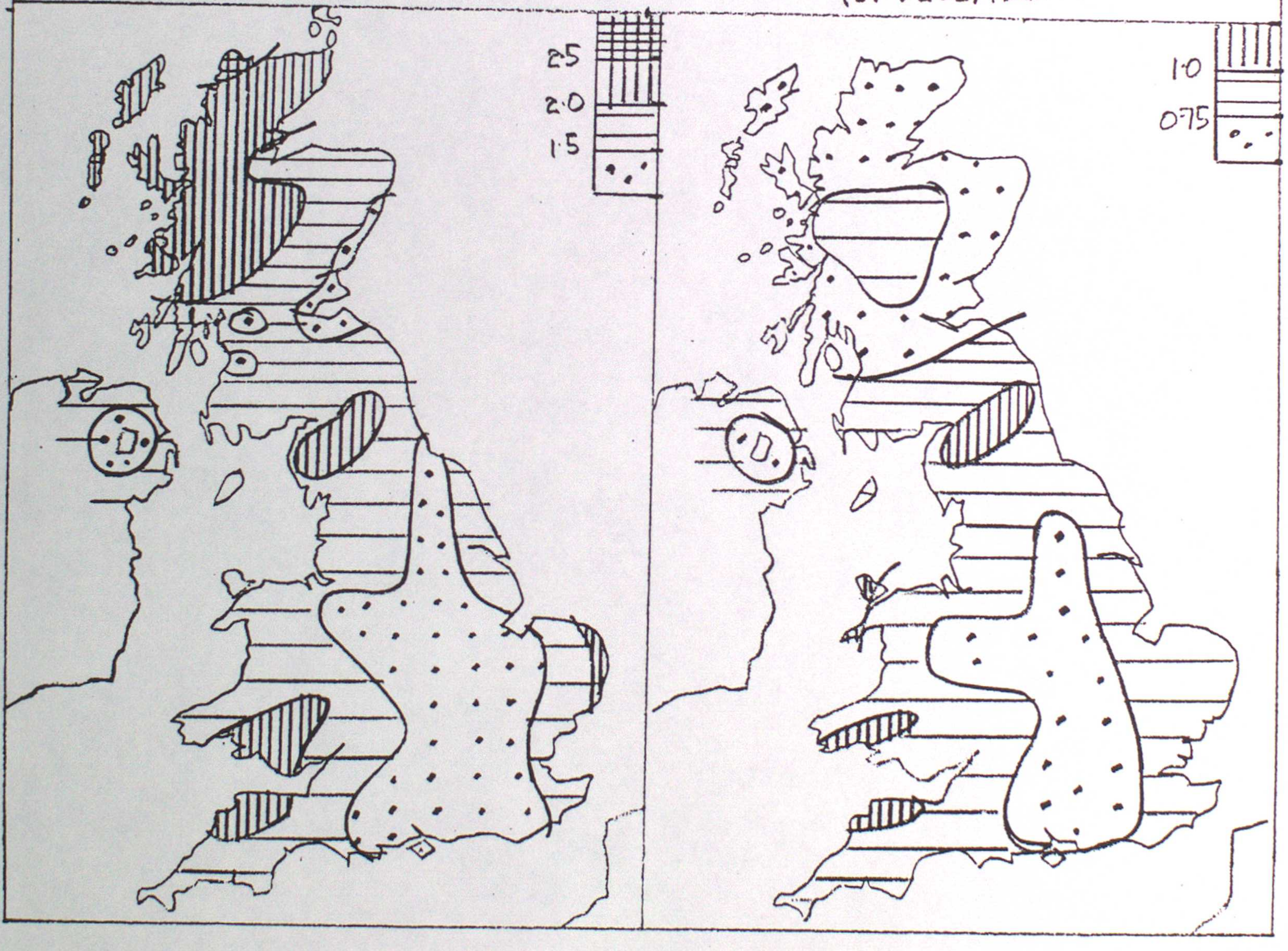


FIG 8 - GEOGRAPHICAL VARIATIONS IN RMS RESIDUALS (HOURS) OF DAILY SUNSHINE OBTAINED FROM PRINCIPAL COMPONENT ANALYSIS (S-MODE, DAILY ANOMALIES, 15 COMPONENTS, MEAN OF 5 ANNUAL ANALYSES IN PERIOD 1979-83)



(a) JUNE

(b) DECEMBER

FIG 9 - GEOGRAPHICAL VARIATIONS IN RMS RESIDUALS OF MINIMUM TEMPERATURE (DEG C) USING NEAR NEIGHBOUR TECHNIQUE (1973 TESTED ON 1978)

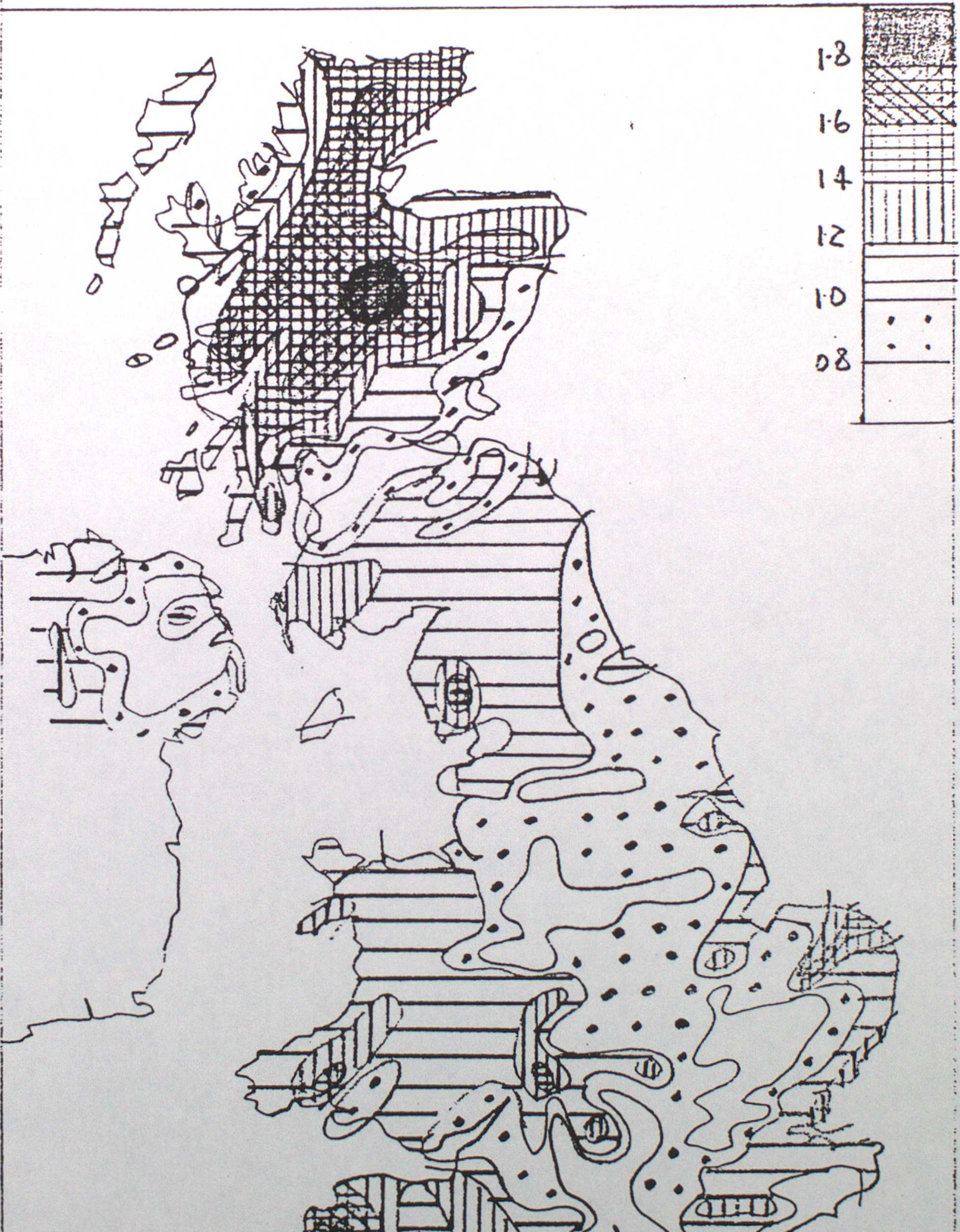
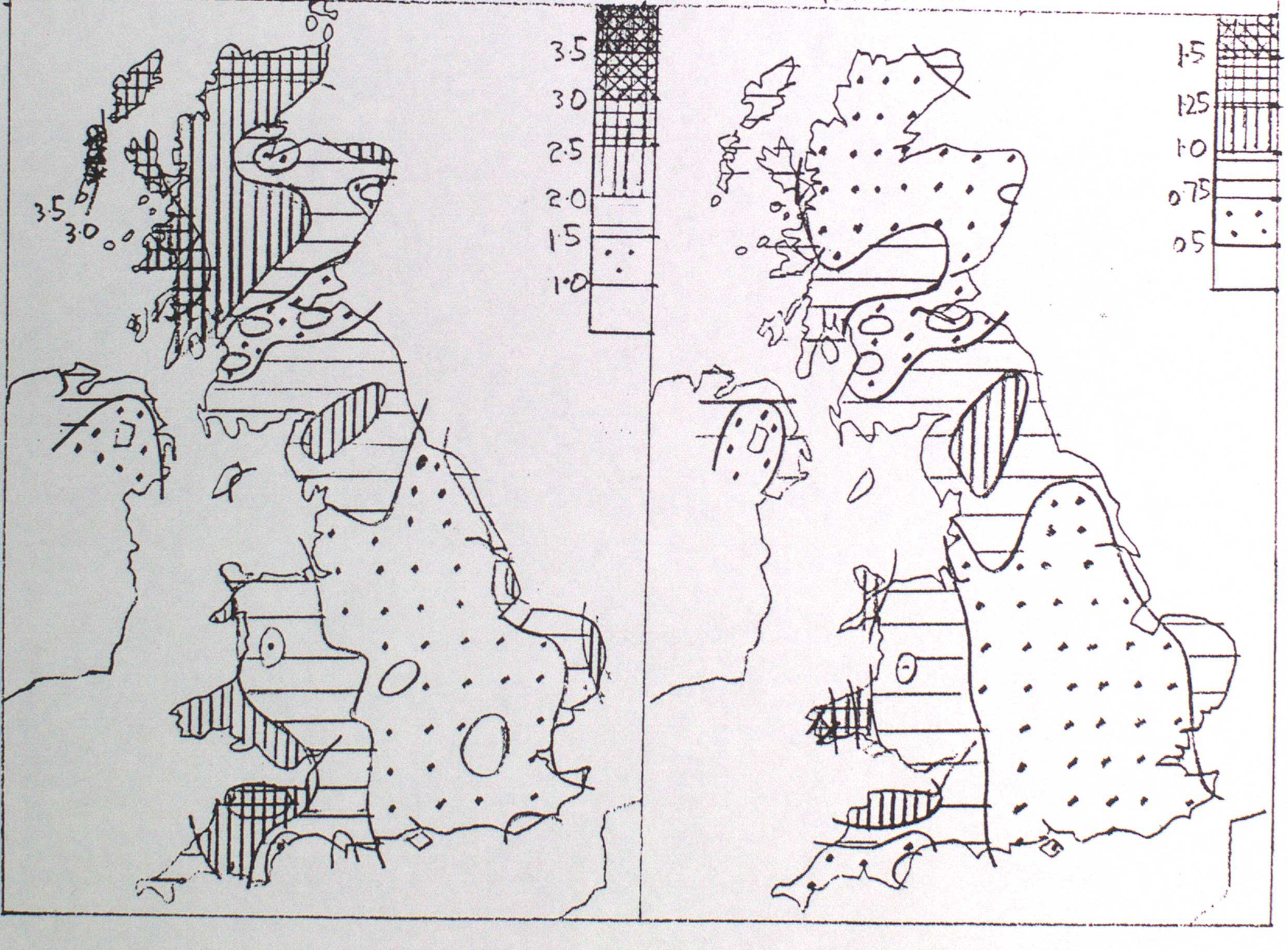


FIG 10 - GEOGRAPHICAL VARIATIONS IN RMS RESIDUALS (HOURS) OF DAILY SUNSHINE USING NEAR NEIGHBOUR TECHNIQUE (MEAN OF 5 ANNUAL ANALYSES IN PERIOD 1979-83)



(a) JUNE

(b) DECEMBER

FIG11 - TUNING OF THE QUALITY CONTROL OF MINIMUM TEMPERATURE (1973 TESTED ON 1978)

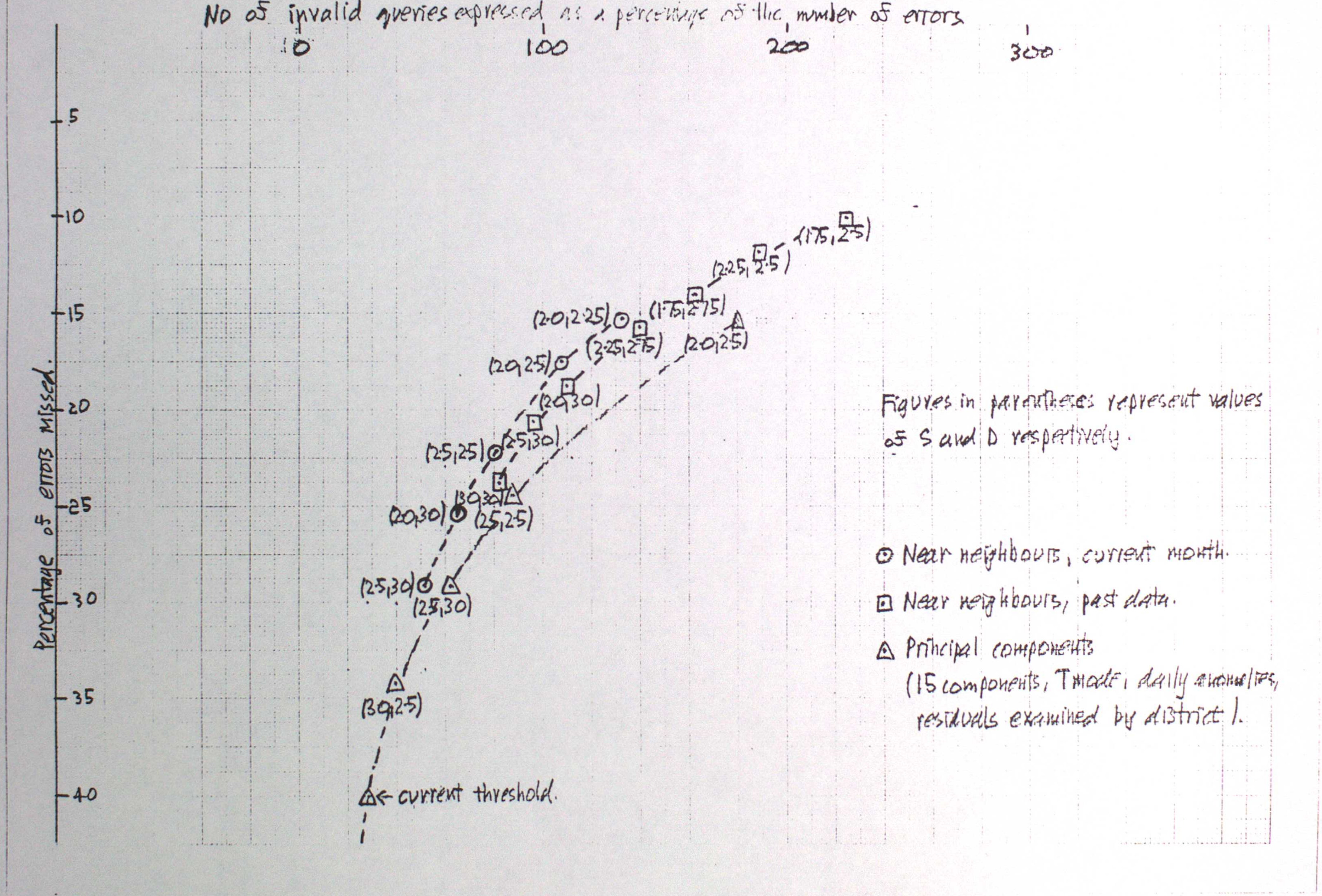


FIG12–TUNING OF THE QUALITY CONTROL OF DAILY SUNSHINE. (1979 TESTED ON 1983)

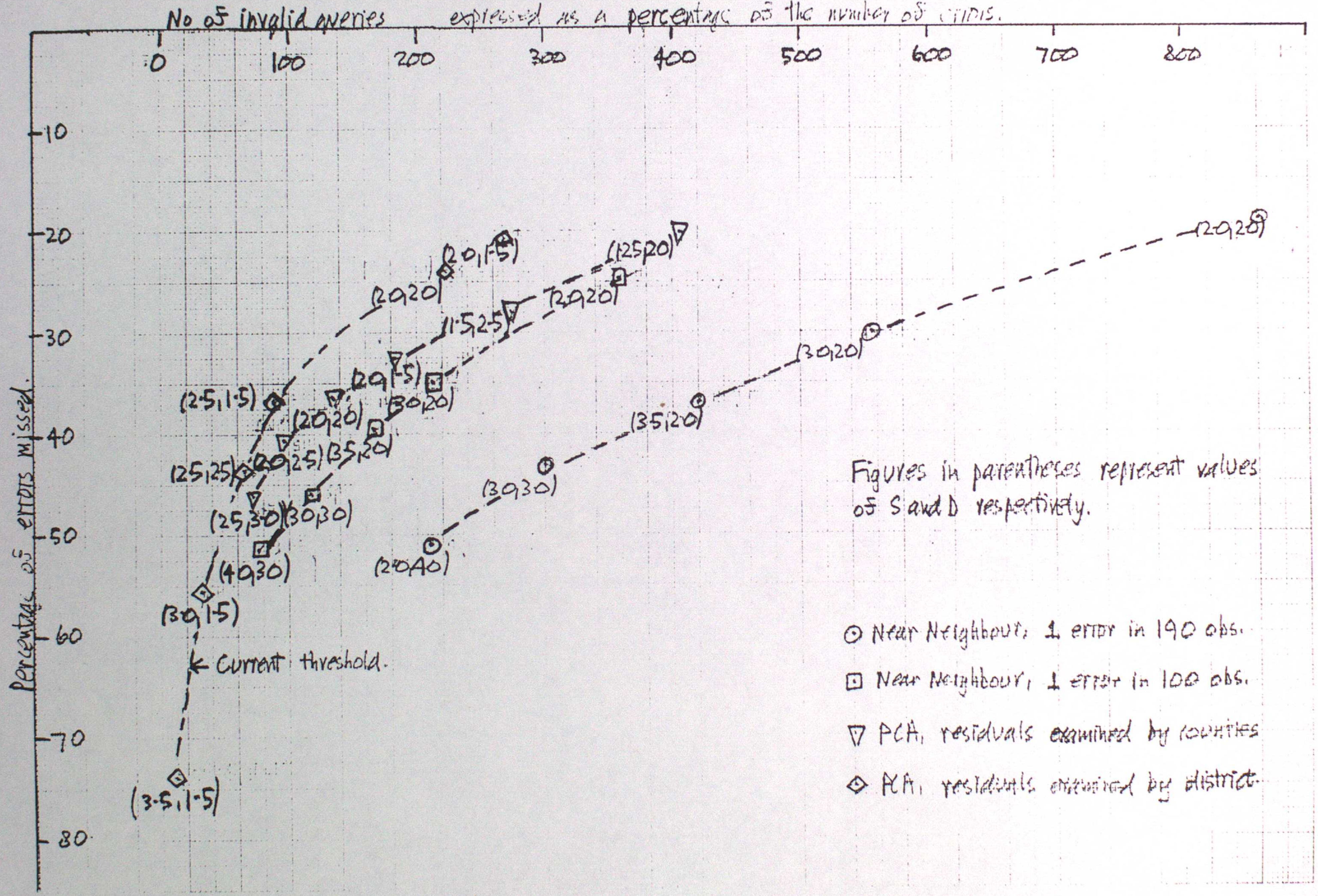
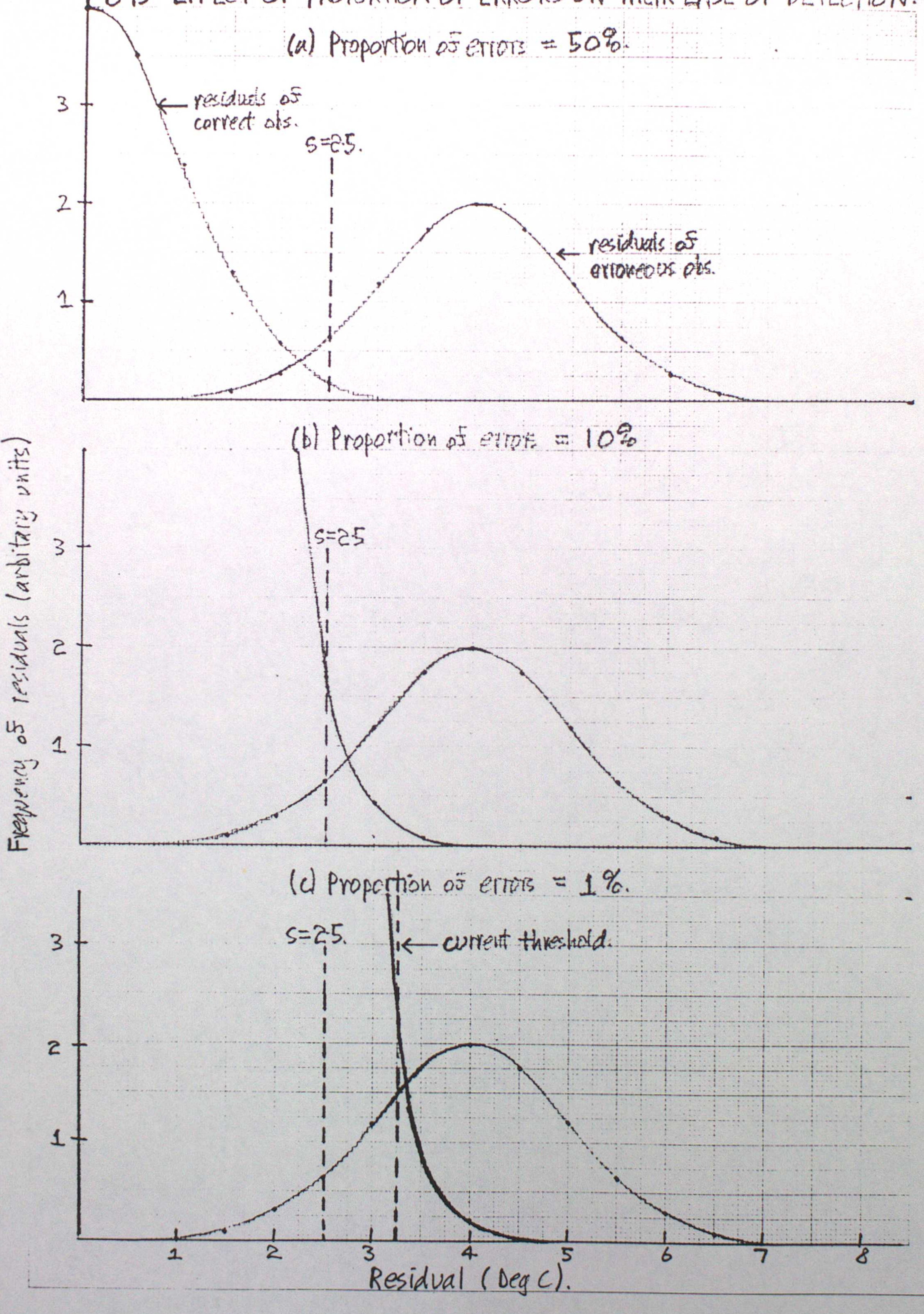


FIG 13 - EFFECT OF PROPORTION OF ERRORS ON THEIR EASE OF DETECTION.



(a) Proportion of errors = 50%.

(b) Proportion of errors = 10%

(c) Proportion of errors = 1%.

Table 1 - Effect of type of anomaly and unit used for input to a principal component analysis of daily sunshine.
S mode analysis of 197 stations in UK for 1979 tested on 1981.

Number of components	RMS residuals (hours) of erroneous observations								
	Station Anomalies			Daily Anomalies			Local Anomalies.		
	Hours	% Poss.	% Ave.	Hours	% Poss.	% Ave.	Hours	% Poss.	% Ave.
Ø	3·31	3·30	3·31	2·40	2·40	2·34	2·35	2·36	2·34
1	2·34	2·34	2·34	2·04	2·03	2·01	1·99	1·98	2·01
2	1·98	1·98	1·99	1·99	1·98	1·92	1·92	1·90	1·92
3	1·89	1·89	1·89	1·84	1·84	1·77	1·76	1·75	1·77
6	1·56	1·55	1·55	1·56	1·56	1·46	1·52	1·51	1·46
9	1·43	1·41	1·40	1·39	1·39	1·38	1·37	1·36	1·38
12	1·38	1·36	1·35	1·34	1·34	1·35	1·35	1·35	1·35
15	1·33	1·33	1·35	1·31	1·28	1·33	1·31	1·28	1·33
18	1·28	1·28	1·34	1·28	1·26	1·33	1·28	1·26	1·33
21	1·29	1·26	1·32	1·28	1·26	1·32	1·28	1·26	1·32

Table 2 - Comparison of regional and national principal component analyses of minimum temperature
Daily anomalies for 1973 tested on 1978 (deg C).

Mean of 10 regional analyses (S-mode)				UK analyses (T-mode)			
	RMS residuals from				RMS residuals from		
	All obs in		Erroneous		All obs in		Erroneous
No of comps	Original Data	Independent Data	obs only	No of comps.	Original Data	Independent Data	obs only.
Ø	1·60	1·50	1·60	Ø	2·36	2·14	2·27
1	1·22	1·17	1·28	1	1·80	1·71	1·79
2	1·08	1·05	1·24	2	1·49	1·45	1·52
3	0·97	0·97	1·22	3	1·39	1·38	1·46
4	0·90	0·92	1·24	6	1·19	1·21	1·28
5	0·85	0·88	1·34	9	1·08	1·14	1·22
6	0·80	0·86	1·45	12	1·00	1·10	1·15
7	0·76	0·84	1·55	15	0·94	1·06	1·13
8	0·72	0·81	1·73	18	0·89	1·04	1·11
9	0·68	0·80	1·90	21	0·84	1·02	1·09
1Ø	0·64	0·78	2·07	24	0·80	1·01	1·09

Table 3 – Comparison of seasonal and annual principal component analyses of daily sunshine

Local anomalies in hours for 197 stations in UK in 1979–80 tested on 1982–83

	RMS residuals (hours) of erroneous observations	
Number of components	Seasonal (T-mode)	Annual (S-mode).
φ	2·39	2·40
1	2·02	1·97
2	1·92	1·81
3	1·81	1·68
6	1·55	1·47
9	1·41	1·32
12	1·35	1·29
15	1·32	1·25
18	1·30	1·25

Table 4 – Effect of number of years data supplied on principal component analysis of minimum temperature.

S-mode analyses of daily anomalies for 189 stations in UK tested on 1978–82 and based on data from

	1975			1973–77		
	RMS residuals (degC) from			RMS residuals (degC) from		
	All observations in		Erroneous	All observations in		Erroneous
Number of components	Original Data	Independent Data	observations only	Original Data	Independent Data	observations only.
φ	2·23	2·16	2·24	2·17	2·16	2·24
1	1·78	1·70	1·72	1·67	1·69	1·71
2	1·49	1·43	1·49	1·42	1·42	1·47
3	1·35	1·34	1·40	1·33	1·33	1·38
6	1·15	1·18	1·16	1·14	1·14	1·12
9	103	1·10	1·11	1·04	1·06	1·07
12	093	1·05	1·11	0·97	0·99	1·04
15	0·86	1·00	1·12	0·92	0·94	1·06
18	0·80	0·97	1·13	0·87	0·91	1·07

Table 5 - Effect of performing a second principal component analysis on minimum temperature

RMS residuals (deg C) of erroneous observations for analyses derived from daily anomalies and tested on data for 1978-82.

Number of	T-mode analyses based on data for					S-mode	Second PCA
components	1973	1974	1975	1976	1977	1973-77	1973-77
φ	2·29	2·29	2·29	2·29	2·29	2·29	2·29
1	1·78	1·78	1·79	1·79	1·78	1·77	1·77
2	1·48	1·48	1·48	1·48	1·47	1·46	1·46
3	1·38	1·38	1·37	1·37	1·38	1·36	1·36
6	1·21	1·22	1·22	1·22	1·20	1·17	1·17
9	1·14	1·16	1·15	1·15	1·12	1·09	1·09
12	1·09	1·11	1·11	1·11	1·10	1·04	1·05
15	1·07	1·09	1·06	1·08	1·07	1·00	1·00
18	1·07	1·08	1·06	1·07	1·05	0·98	0·98

Table 6 - Comparison of single and multi-element analyses of temperature.
T-mode analyses of daily anomalies for 1973 tested on 1978

RMS residuals (deg C) of erroneous observations.

	Minimum		Maximum		Dry Bulb		Wet Bulb.	
No of comps	Single Element	Multi-Element	Single Element	Multi-Element	Single Element	Multi-Element	Single Element	Multi-Element
φ	2·17	2·17	2·02	2·02	2·15	1·96	1·86	1·77
3	1·37	1·54	1·30	1·39	1·43	1·29	1·13	1·13
6	1·24	1·44	1·12	1·31	1·25	1·22	0·97	1·04
9	1·19	1·44	1·02	1·24	1·15	1·16	0·92	0·99
12	1·15	1·34	0·96	1·14	1·12	1·10	0·91	0·96
15	1·12	1·27	0·94	1·11	1·09	1·07	0·90	0·96
18	1·11	1·25	0·93	1·08	1·10	1·05	0·92	0·94
21	1·09	1·25	0·93	1·07	1·09	1·04	0·91	0·93
24	1·10	1·24		1·07	1·07	1·02		0·94
27		1·21		1·06		1·02		0·94
30		1·18		1·06		1·01		0·93
33		1·17		1·05		0·99		0·93
36		1·18		1·05		0·98		0·91
39		1·17		1·03		0·98		0·91

Table 7 - RMS residuals associated with near neighbour technique based on past data.

Regression Technique	Data used	Treatment of seasonal variation	Min Temp (deg C)	Sunshine (hours)
Constant Difference/Ratio	30 year averages	Mean of all months	1·00	1·10
"	"	Monthly smoothing	0·99	1·09
"	1 year calibration period	Mean of all months	1·02	1·10
"	"	Monthly smoothing	1·02	1·10
"	4 year calibration period	"	0·95	1·08
None (station = neighbour)	Current month.	None	1·16	1·11

1 year calibration period = 1973 for min temp and 1979 for sunshine
4 year " " = 1973-76 " 1979-82 "
Test period = 1978 " 1983 " .

Table 8 - RMS residuals associated with near neighbour technique using data from current month.

Regression Technique	Data used for		Min temp (deg C)		Sunshine (hours)	
	Selection of neighbours	Weighting of neighbours	All obs	Erroneous obs	All obs	Erroneous obs.
Constant Diff/Ratio	1 year calibration period	1 year calibration period	0·90	1·03	1·05	1·10
Slope = ratio of SD's.	"	"	0·83	1·04	—	—
Constant Diff/Ratio	Current month	1 year calibration period	0·81	1·08	1·01	1·16
Slope = ratio of SD's.	"	"	0·74	1·10	—	—
Constant Diff/Ratio	Current month	Current month	0·78	1·13	1·00	1·19
Slope = ratio of SD's.	"	"	0·71	1·17	—	—

Calibration period = 1973 for minimum temperature and 1979 for sunshine.
Current month taken from 1978 " 1983 " .

Table 9 - RMS residuals of minimum temperature for new stations using the principal component technique and data from current month.
Component scores estimated by BMDPAM using ALLVALUE option.
T-mode analyses of daily anomalies. in deg C.
Mean of 5 annual analyses for 1973-77 tested on 1978-82.

	Erroneous observations						All observations					
No of comps	Real scores	Scores estimated with RIDGE=					Real scores	Scores estimated with RIDGE=				
		1·0	0·9	0·7	0·5	0·3		1·0	0·9	0·7	0·5	0·3
φ	2·22	2·23	2·23	2·23	2·23	2·23	2·16	2·19	2·19	2·19	2·19	2·19
1	1·74	1·84	1·82	1·79	1·79	1·80	1·71	1·71	1·70	1·70	1·70	1·73
2	1·48	1·46	1·42	1·41	1·43	1·51	1·45	1·37	1·37	1·37	1·41	1·49
3	1·41	1·46	1·42	1·39	1·40	1·47	1·37	1·27	1·27	1·29	1·34	1·44
6	1·25	1·58	1·49	1·41	1·39	1·44	1·21	1·07	1·08	1·14	1·22	1·36
9	1·19	1·73	1·60	1·46	1·41	1·44	1·14	0·99	1·01	1·08	1·18	1·33
12	1·16	1·88	1·71	1·52	1·42	1·44	1·09	0·93	0·96	1·04	1·15	1·32
15	1·13	1·97	1·78	1·55	1·44	1·44	1·06	0·88	0·93	1·02	1·13	1·31

Table 10 - Performance Scores of Quality Control of Minimum Temperature using Principal Components.
T-mode analysis of daily anomalies ~~by~~ for 1973 tested on 1978

Treatment of residuals			Flagging Threshold		No of components					
Flagging by	Mean set to zero	trimmed	D (deg C)	S	5	10	15	20	30	40
District	✓	x	2·5	2·5	39·6	32·8	32·1	32·4	31·9	33·6
District	✓	✓	2·5	3·5	38·9	33·6	32·9	32·3	31·5	33·2
District	x	x	2·5	2·5	35·3	32·7	32·7	31·7	32·5	32·9
District	x	✓	2·5	3·5	34·1	34·9	33·3	32·6	32·1	32·2
Counties	✓	x	2·5	2·0	39·9	33·9	32·1	32·5	33·4	33·0
Counties	✓	✓	2·5	3·0	39·8	34·9	32·7	33·3	33·1	32·6
Counties	x	x	2·5	2·0	33·0	31·6	30·2	30·7	32·6	33·2
Counties	x	✓	2·0	3·0	32·9	33·1	32·2	32·3	32·7	32·8

Table 11 - Performance Scores of Quality Control of Daily Sunshine using Principal Components.
S-mode analysis of daily anomalies for 1979 tested on 1983.

Treatment of residuals			Flagging Threshold		No of components.				
Flagging by	Mean set to zero	trimmed	D (hours)	S	5	10	15	20	25
District	✓	x	1·5	2·5	49·6	47·9	44·8	44·7	42·3
District	✓	✓	1·5	4·0	48·2	45·1	42·4	41·9	41·9
District	x	x	1·5	2·5	48·1	47·6	45·7	44·9	43·4
District	x	✓	1·5	4·0	46·4	46·3	43·3	42·2	41·8
Counties	✓	x	2·0	2·0	56·1	52·8	48·8	48·2	47·0
Counties	✓	✓	2·5	3·0	61·3	55·5	52·7	49·7	48·0
Counties	x	x	2·0	2·0	50·2	50·7	49·9	49·4	48·5
Counties	x	✓	3·0	3·0	44·7	49·4	48·8	48·0	46·8

Table 12 - Effect of varying the frequency of errors on a principal component analysis of daily sunshine.

Daily anomalies in hours for 1979 tested on 1983 for 197 stations in UK

Number of components	RMS residuals (hours) of erroneous observations for errors inserted with the following frequencies (1 in X obs).				
	10	30	100	190	365.
ϕ	2·58	2·59	2·51	2·60	2·63
3	1·85	1·81	1·86	1·83	1·76
6	1·61	1·56	1·59	1·58	1·53
9	1·48	1·42	1·44	1·38	1·38
12	1·42	1·35	1·39	1·34	1·31
15	1·39	1·31	1·31	1·27	1·31
18	1·38	1·31	1·31	1·29	1·26
21	1·39	1·32	1·31	1·29	1·23
24	1·42	1·34	1·32	1·32	1·27