LONDON, METEOROLOGICAL OFFICE.

Met.O.3 Technical Note No.8.

The estimation of missing climatological data. By TABONY,R.C.

London,Met.Off.,Met.O.3 Tech.Note No.8, 1981,31cm.Pp.25,11pls.

Met O 3 Technical Note No 8

THE ESTIMATION OF MISSING CLIMATOLOGICAL DATA

by

R C Tabony

Climatological Services Branch (Met O 3)
Meteorological Office
London Road
Bracknell
Berkshire
RG12 2SZ

October 1981

CONTENTS

Summary

Various methods of estimating monthly means and extremes of climatological data are examined. Any generalised method is likely to be based on a correlation matrix, but the incompleteness of the data introduces problems with this approach. These are illustrated by program BMDPAM of the BMDP suite, which produces estimates worse than those using traditional methods based on single station comparisons. Principal component analysis, (a technique not included in BMDPAM), is likely to be the best statistical tool for estimating missing values among highly correlated data. It requires, however a higer quality correlation matrix than is normally attainable from incomplete data. This can be obtained by using a simple estimating procedure to produce a preliminary set of complete data. A simple technique was devised for estimating climatological data in the UK, and this was found to give results similar to those obtained from an eigenvector scheme used for quality control purposes. Further refinement of the estimates therefore seemed unnecessary. The accuracy of the estimates is such that it is suggested that satisfactory averages could be computed from only 10 years of data, and possibly less.

1. Introduction

Most meteorological investigational work involves the analysis of data and these are very rarely complete. The problem of missing data is therefore very common and a simple and effective procedure for eliminating it would be of considerable practical value. In climatology, one of the most obvious applications of such a procedure is in the production of averages and other statistics. Their publication is one of the most important functions of National Meteorological Services, and they are put to a wide variety of uses by a large section of the community.

The period of data to be used in the compilation of climatological averages has been the subject of much discussion (eg Jagannathan et al, 1967). The final choice is essentially a compromise between a short period, which gives rise to sampling errors, and a long period, which fails to keep pace with climatic change. Arguably, therefore, the optimum averaging period is obtained when the sampling error becomes of the same order as the likely climatic change. This varies, however, with element, region, and season. Another factor to be taken into account is the length

of homogeneous records commonly available at single sites. The shorter the period that can be used to provide satisfactory averages, the greater the number of stations that can be used to answer climatological enquiries. As a compromise between conflicting interests, WMO recommend a universal averaging period of 30 years.

In the UK, the amount of missing data that has to be estimated in the calculation of 30 year averages is considerable. In the period 1951-80, for instance, over 600 stations recorded temperature for 10 or more years, yet the number with complete records in the 30 year period is only 50. It is clear that if a reliable method of estimating missing data were available, then the number of stations for which 30 year averages could be produced would be considerably increased.

This paper is primarily concerned with averages of monthly means and extremes. The best method of estimating these could well be the examination of daily data, but this only considers methods based on the monthly means or extremes themselves. Daily data were not used since this would greatly increase the magnitude of the task and, as will be shown, adequate estimates can be obtained using monthly values. The only elements considered are temperature and sunshine, although the discussion may reasonably be extended to other climatological parameters.

30 or 35 year averages of temperature and sunshine have been produced by the UK Meteorological Office fairly regularly since 1900 and, during this time, the method of estimating missing data has remained unchanged. It is a simple procedure based on comparisons with a single neighbour, the required computations being performed manually. A change to a more sophisticated computer-based technique is clearly desirable. The BMDP suite of statistical software described by Dixon and Brown (1979) contains a program BMDPAM which was designed to estimate missing data. This contains a number of options which were thoroughly tested, and all proved unsatisfactory. A simple routine based on linear regression was then written and tested.

2. Data

Since 30 years is the recommended length of an averaging period, it would clearly be sensible to use 30 years of data for this investigation. At the time of writing, data from the entire UK climatological network are only held in convenient computer form for the years 1959-1979 and this dictated the period of data used. The elements examined were:-

(i) Maximum and minimum temperatures observed at 09h every morning, and averaged over all days in a particular month. These are referred to as the 'daily max' and 'daily min' respectively.

(ii) The highest maximum and lowest minimum temperatures recorded in each month. These are referred to as the 'monthly max' and 'monthly min' respectively. From these data series the highest monthly max and the lowest monthly min in a period of years are referred to as the 'extreme max' and 'extreme min' respectively.

(iii) The total number of hours of bright sunshine recorded in a month, referred to as the 'monthly sun'.

The largest sources of error in station averages are likely to be caused by inhomogeneities due to sudden or gradual changes of site or instrumentation. Ideally, these should be eliminated before comparisons between neighbouring stations are made. In practice, however, this is a difficult and time-consuming task; and the investigations which follow are based on data from which inhomogeneities have not been removed. The data had, however, been subjected to simple quality control procedures which removed climatologically 'impossible' values.

The stations chosen were those with 120 months or more of data present in the period 1959-79, plus 23 selected stations for temperature and 10 for sunshine. This gave networks of around 570 stations for temperature and 370 for sunshine, and their distributions are displayed in figs 1 and 2 respectively. For temperature the station density is reasonably uniform apart from northwest Scotland, where it is much lower than elsewhere. Far fewer inland stations report sunshine than temperature but the majority of stations along the coast observe both temperature and sunshine.

3. Background climatology

The best methods of estimating missing data will, in general, depend upon the statistical properties of the data. In climatology, the two most important factors are the inter-correlations in the station network, and the seasonal variations in the relations between stations. Some idea of these variations in the UK is gained by examining values obtained for 10 stations - Achnashellach (isolated), Braemar (frost hollow), Durham (standard site), Santon Downham (frost hollow), Oxford (standard site),

Hastings (coastal), Earls Hill (upland), Corwen (frost hollow), Scilly (all neighbours in one direction), and Armagh (standard site). For sunshine, Achnashellach was replaced by Fort Augustus and Corwen by Douglas. The locations of these stations are shown in fig. 3.

The highest correlations between neighbouring stations for temperature and sunshine are shown in fig. 4. The figures represent the average of results obtained from the 10 stations in the period 1959-79. Not surprisingly, the highest correlations, around 0.98 to 0.99, are observed for daily max, but there is some seasonal variation, with higher values in winter than in summer. Daily minima show a more pronounced seasonal variation, with correlations ranging from 0.95 in summer to 0.98 in winter. Hence daily min in winter is as highly correlated as daily max in summer. Monthly max and monthly sun both give correlations between neighbours around 0.95, with little seasonal variation (but remember there is a less dense network of stations for sunshine than for temperature). As expected, the lowest correlations between neighbours are observed for monthly min, with values ranging from around 0.8 in summer to 0.9 in winter.

Some idea of the decay of correlation in the station network is given in table 1, where associations between neighbours with the highest and 6th highest correlations with one another are compared. Averaged over the 5 parameters and 10 stations examined, the correlation falls from 0.94 for the first ranked neighbours to 0.90, for the 6th ranked neighbours. For daily max at places like Oxford, where there is a relatively dense network of stations set in topographically simple country, the 6th ranked neighbour has a correlation as high as 0.99. For monthly max at places like Scilly, however, where all neighbours lie in the same direction and are relatively distant, the correlation of the sixth neighbour has fallen to 0.71.

The seasonal variations in the standard deviation of temperature and sunshine are illustrated in fig. 5. For sunshine, the standard deviation varies roughly in proportion to the mean sunshine received. For temperature, the variations are very similar to those observed for the correlations in fig 4. Thus a decrease in the correlation, which makes estimates more difficult to make, is accompanied by a decrease in the standard deviation, which renders the estimates more reliable.

A more detailed examination of neighbouring stations is made in fig 6, which illustrates seasonal variations in

(i) the correlation between the stations

(ii) the difference in the monthly means (Δ)

(iii) the ratio of the standard deviations, which corresponds to the slope of a straight-line relation between the stations.

Items (ii) and (iii) may be regarded as the coefficients in a linear relationship between two stations. Fig 6 shows that there can be considerable seasonal variations in these coefficients.

4. Review of techniques

4.1 Problems associated with generalised methods based on the correlation matrix

A general approach to the estimation of missing data will commonly start from the correlation matrix. This is usually required to be complete and self-consistent, and this introduces the following difficulties:-

(i) If two stations records do not overlap, then one would not normally be used to make estimates for the other. If the correlation matrix is required to be complete, however, an estimate must be made, and this may not be very good. Because the data is incomplete the correlation matrix is not self-consistent, but can be made to be so by 'smoothing' of the elements of the correlation matrix. In this procedure, any errors in the 'missing' elements are shared amongst the others, degrading any estimates which may be derived from them.

(ii) Selection of the 'best' neighbours should depend not only on the inter-correlations between stations, but also on the length of overlap between them. If two neighbours had the same correlation with a station, for example, then the preferred neighbour would clearly be that with the longer overlap. Furthermore, correlations based on short overlaps have a larger random error than those based on long overlaps, so the highest values in a correlation matrix are much more likely to be derived from short overlaps than long ones. When using a correlation matrix, therefore, it is important to be able to weight the stations according to the length of overlap with one another.

(iii) In some statistical packages, the correlation matrix is constructed from standardised anomalies based on means and standard deviations derived from all available data for each station. The mean and standard deviation for each station are therefore based on disparate periods. In general, the correlations calculated in this way will be too low. Suppose, for instance, that station A recorded observations for a period in which cold years predominated, while station B operated during a period in which the majority of years were warm. Any years with near average temperatures would then be credited with a positive anomaly at station A and a negative anomaly at station B. Clearly the true correlation between two stations can only be estimated from data which is restricted to the period of overlap between the stations.

(iv) The data from which the correlation matrix is derived must belong to a single population. This is obvious, of course, and in most applications imposes no restrictions. In climatology, however, the seasonal variation means that each month or season must be treated separately, thereby restricting the data sample that can be supplied. A 'mixed season' analysis could be made by expressing the data in the form of standardised anomalies, but the monthly means and variances would be based on small sample sizes and disparate periods, so that problems remain.

4.2 Minimisation of RMS errors and loss of variance

If observations are standardised, the gradient of a least squares linear regression between two variables is equal to the correlation between them. Thus if two variables had a one to one relationship with one another, but there was a random scatter in the observations such that the correlation between them was only 0.7, then the slope of the linear regression would also be 0.7. Hence minimising RMS errors generally leads to a loss of variance. If there are not many missing values, or they are only going to be used to estimate means, this may be quite acceptable. In general, however, the more of the original variance that can be retained, the better.

4.3 Total v Partial Correlation

One of the simplest ways of estimating missing observations is to express them as a linear combination of observations from neighbouring stations. A basic problem is to decide whether the weights associated with the neighbours should be determined on the basis of partial or total correlation. The partial correlation between a station and a neighbour is the correlation between them after the effects of other neighbours have been removed. It is the basis of multiple linear and stepwise regression techniques. In multiple linear regression, all the neighbours supplied are included in the prediction equation. In stepwise regression, neighbours are used as predictors only if their partial correlation with the test station is significantly different from zero.

An observation at a given station may be expressed as

linear relation x observation at a neighbour + systematic (geographical) difference + random difference.

The linear relation with the first neighbour will be taken into account in both the partial and total correlation approaches. The total correlation method reduces random errors but neglects spatial changes. Stepwise regression takes into account geographical variations but neglects random differences. The behaviour of multiple linear regression depends on the inter-correlations between the variables supplied; if these are low, the results will be similar to those obtained from stepwise regression, while if they are high, they will be similar to those derived from total correlations.

In section 3.1 it was shown that, for the stations and elements examined, the highest correlations between stations averages 0.94. This value is sufficiently high that the use of a stepwise regression technique would lead to only one neighbour being retained in the prediction equation. In these circumstances multiple linear regression might be expected to provide better estimates. When the data are highly correlated, however, the equations used to determine the regression coefficients (weights) associated with each station are ill-conditioned, and the coefficients produced have large standard errors. Ridge regression is a technique designed to over-

come this problem. It is equivalent to dividing the off-diagonal elements of a correlation matrix by 1+K, where K is known as the ridge parameter. A good general description of ridge regression is given by Meisner (1979). It is one of many techniques made available by program BMDPAM.

To illustrate the use of ridge regression, consider the case of 10 neighbours whose correlations with a test station all have expected values of 0.9. In ordinary least squares regression, the station weightings have expected values of 0.1, but with large standard errors, also 0.1 (say). Application of ridge regression might lead to expected weightings of 0.08, but with much smaller standard errors, 0.01 (say). It can be seen that the stability of the regression coefficients is obtained at the expense of a loss of variance in the final estimates.

When the inter-correlations in the station network are as high as for climatological stations in the UK, differences between neighbours due to large scale geographical variations are small. In this situation, beighbours selected on the basis of their total correlation with a test station are likely to yield better estimates than those based on partial correlation; stepwise regression is likely to produce estimates based on only one neighbour, while a combination of multiple linear and ridge regression leads to loss of variance.

4.4 Optimum and linear interpoloation

The term optimum interpolation is generally used to describe a technique in which estimates at a point are derived from a linear conbination of observations at neighbouring stations, the station weights being chosen to minimise the RMS errors of the estimates. It is therefore similar to multiple linear regression. The term linear interpolation is usually used to describe a method in which estimates are obtained by fitting a plane to the surrounding stations.

Linear interpolation gives results which are only slightly inferior to optimal interpolation if

(i) the stations are reasonably uniformly distributed

(ii) the correlation between nearest neighbours is high, and between points close together very high (> 0.9)

(iii) the correlation decay function is homogeneous and isotropic (ie independent of location and direction)

Hopkins (1977) applied linear interpolation to temperature and sunshine data in East Anglia, where the uniform terrain enabled the above conditions to be most nearly satisfied.

Neither linear nor optimum interpolation was tried in this paper. Linear interpolation is unable to cope with the effects of topography, while optimum interpolation, being based on a correlation matrix, suffers from the pitfalls mentioned in section 4.1. The RMS errors obtained by the method proposed in this paper are, however, compared with those obtained by Hopkins (1977) in section 8.

4.5 Principal component analysis

Principal component analysis is fully described by Kendall (1975). It enables fields of correlated data to be represented by a set of orthogonal patterns or eigenvectors, each of which explains the greatest part of the (remaining) variance. The leading eigenvectors represent systematic differences between the stations while the random differences are consigned to higher order components. By reconstituting the data from only the leading eigenvectors, therefore, the genuine differences between stations are retained while the noise is ignored. As thus described, the technique seems an ideal means of estimating missing values. The disadvantage is that a complete and self-consistent correlation matrix is required so, if the data are incomplete, the problems outlined in section 4.1 are encountered.

Principal component analysis forms the basis of the routines currently operated by the UK Meteorological Office for the quality control of daily climatological data (Spackman, 1980). Spackman achieved a complete and self-consistent correlation matrix by using a simple estimating procedure to obtain complete data prior to performing the principal component analysis. He eliminated missing values by taking the mean of all observations in the same county for the day in question and adjusted it by the annual average difference between station and county values. The same approach could be used in the estimation of missing data. A simple method could be used to make 'first guess' estimates which could be refined by a subsequent principal component analysis. A suitable technique based on the concept of total correlation was developed and is described in section 7. The comparison of results in section 8, however, shows that the estimates obtained using this method are of similar quality to those achieved by the quality control routines. Subsequent refinement of these estimates by principal component analysis therefore seemed unnecessary.

4.6 Traditional and proposed techniques

The method traditionally used by the UK Meteorological Office to estimate missing temperature and sunshine data is based on comparisons with a single neighbouring station. For temperature, a constant difference between stations has been assumed. Thus if the January temperature at station A has been 0.1°C above that at station B during a period of overlapping records, then 0.1°C was added to the values at station B to give estimated values at station A. For sunshine a constant ratio between stations has been assumed. Thus if the July sunshine at station A has been 1% less than at station B during a period of overlapping records, then 1% was subtracted from the values at station B to provide estimated values at station A.

In the proposed technique, the constant difference or ratio is replaced by a linear regression, and the single neighbour is replaced by several. The main points of the method are

(i) To select the best neighbours, the (total) correlations with the test station are averaged over all months, and adjustments made according to the length of overlap.

(ii) To preserve variance, the slope of the linear regression between two stations is made equal to the ratio of the standard deviations (ie the moderating effect of correlation is ignored).

(iii) To ensure a smooth seasonal variation, the regression coefficients are smoothed over a number of months.

(iv) To reduce random errors, the final estimate is based on a linear combination of estimates obtained from individual neighbours.

5. Estimation using the traditional method and evaluation of techniques

A computer program was written to simulate the procedures involved in the traditional method of estimating temperature (which in practice were carried out by hand). The neighbour selected was that for which the standard deviation of temperature difference with the test station was smallest. Any seasonal variation in the difference was retained in order to prevent stations of different site characteristics (eg coastal, inland) from being matched with one another. Only neighbours which were capable of eliminating all missing values were considered. A similar routine, based on ratios rather than differences, was written for sunshine.

All the methods of estimating missing data which are examined in this paper were evaluated as follows. An array of data, consisting of points in space (stations) by points in time (years) was formed. A station with complete or almost complete records was selected as test station, and its observations were withheld for a pre-determined set of years. The estimates produced for the test station for the 'missing' years were then compared with the observations.

The traditional method was evaluated by using the same 10 stations as were used in section 3. Observations for these stations were supplied in turn for the years 1959-63, 1975-79, 1959-68, 1970-79, 1959-73 and 1965-79, and results obtained from this combination of 10 stations and 6 overlapping periods are presented in table 2. The elements examined are monthly and daily min in January, monthly and daily max in July, and monthly sun in June and December. In practice, the method was only used to provide averages of daily max and min and monthly sun, but the RMS errors of the individual estimates are given in table 2 for completeness. Comparison with fig 5 shows that this very simple technique gives RMS errors which are about 40% of the corresponding standard deviations.

6. Estimation using BMDPAM

6.1 The options

The main options are

(i) SINGLE. Linear regression against the most highly correlated neighbour for which a value is present.

(ii) REGR. Multiple linear regression.

(iii) STEP. Stepwise regression.

(iv) TWOSTEP. Stepwise regression limited to a maximum of 2 neighbours.

All the options may be combined with ridge regression, while the level of significance used to retain or reject neighbours in stepwise regression may also be varied. This is achieved by use of an 'F to enter' criterion in which F is defined as

$$F = \left(\frac{\text{slope of regression line}}{\text{standard error of slope}} \right)^2$$

The default value is 4, which corresponds to a significance level of around 5%.

6.2 Presentation of results from 1979 version

The 1979 version of the program BMDPAM was tested on data for Corwen in North Wales from 1959 to 1979. Corwen is a frost hollow and was chosen because its monthly minima in January should be difficult to estimate. This circumstance should afford maximum opportunity for the best estimating procedure to distinguish itself from the remainder (in this kind of situation). For completeness the program was also tested on daily minima in January and daily and monthly maxima in July. The program was supplied with data for 50 stations in Wales and Cheshire in which observations from the test station were restricted in turn to the years 1959-63, 1975-79, 1959-68, 1970-79, 1959-73 and 1965-79. The options examined were SINGLE, TWOSTEP, REGR and STEP for a variety of ridge parameters, and STEP for a range of F to enter criteria. Statistics of the errors of the estimates for monthly minima in January are presented in fig 7. These represent the average over all 6 periods for which Corwen data were supplied.

The behaviour of all 3 statistics displayed in fig 7, namely the RMS error, the error in the average, and the loss of variance, are very similar. The estimates are all very poor, and worse than those obtained from the traditional method. A typical RMS error of 3.3°C may be compared with 2.9°C using the traditional method, while the reduction in the standard deviation of around 1.7°C is considerable (traditional method 0.9°C). The very poor estimates for REGR with a ridge parameter of 0.2 are due to instabilities being generated when 15 years of data from the test station were supplied. The simplest option, SINGLE, performs as well as any, and estimates based on 15 years of data from Corwen are no better than those based on only 5 years. Similar patterns of results emerge when daily minima in January and monthly and daily maxima in July are considered.

6.3 Interpretation of results from 1979 version

The estimates obtained from BMDPAM were poor mainly because they were derived from a correlation matrix, and suffered from the problems described in section 4.1. First consider SINGLE. It has two major faults:-

(i) In selecting the best neighbour, it takes no account of the length of overlap between stations. When the program is supplied with a large number (49) of neighbours, it is almost inevitable that the station selected as best neighbour will be one with a short overlap. This is why the

estimates obtained when 15 years of Corwen data are supplied do not represent an improvement on those obtained from only 5 years of data.

(ii) If two stations do not overlap, the correlation between them is assigned a default value of zero. This introduces inconsistencies into the correlation matrix, and when these are smoothed out, a considerable lowering of the genuine correlations takes place. It will be recalled that when observations are standardised, the gradient of the linear regression is equal to the correlation between them. BMDPAM makes use of this relation, and the reduced value of the correlation is reflected in a decreased slope of the regression line. It is this procedure which leads to the loss of variance which is such a feature of the estimates produced by BMDPAM.

Next consider STEP and TWOSTEP. For the UK climatological network, the first neighbour would commonly account for such a large proportion of the variance that none of the other variables would have partial correlations which were significantly different from zero. In BMDPAM, however, the fit of the first neighbour is not good, and there is more residual variance than there ought to be. In addition, the F to enter criterion (by which neighbours are accepted or rejected) is not adjusted to take account of the amount of missing data. Thus some neighbours with high partial correlations (arising by chance from the brevity of the overlaps) are erroneously accepted as predictor variables.

The erroneous inclusion of neighbours can be prevented by increasing the F to enter criterion or introducing a ridge parameter, which cause the method to revert towards SINGLE. If the F of RIDGE parameters are set high enough, even the first station will fail to satisfy the stringent significance requirements, and the estimates produced are then set equal to the mean of the sample. With the data employed in this investigation, this occurred for a ridge parameter of 2 or more.

6.4 Maximum likelihood estimation of missing correlations (1981 version)

A major fault in the 1979 version of BMDPAM was the setting to zero of the correlations between non-overlapping variables. In the 1981 release (Dixon, 1981), the problem is overcome by the use of a maximum likelihood routine to estimate the missing correlations. This procedure is capable of effecting great improvements in the quality of correlation matrices produced from incomplete data. In BMDPAM, however, the correlations are calculated using means and standard deviations based on all available data for each station, and as described in section 4.1, this results in correlations which are too low. Thus although the maximum likelihood routine is capable of producing estimates of missing correlations which are of the same general level as those present, in BMDPAM its effectiveness in producing unbiased values is compromised by the procedure used to calculate the correlations.

An error in the program resulted in the incorrect assignment of the most highly correlated variable. This has prevented a fair assessment of the effectiveness of the 1981 version of the program from being made.

7. Estimation using proposed technique

7.1 Factors to be considered

The best estimates of missing data can be made by making full use of the available data, ie by using more than one neighbour, and by not treating each month separately. The first point can be satisfied by using a linear combination of neighbours, while the second can be accommodated by some smoothing of the relationships obtained for each month. First consider how the linear combination

of stations should be chosen:-

(i) Each climatological parameter can be estimated separately. The best combination of stations for one element (eg maximum temperature) is not necessarily the same as for another (eg minimum temperature). For maximum temperature the 'best' neighbour is probably the nearest one, but if that nearest neighbour is a frost hollow (and the test station is not), then the best neighbour for minimum temperature will probably be another station which, although further in distance from the original station, is closer in site characteristics.

(ii) The neighbours can be selected according to their total (as opposed to partial) correlation with the test station.

(iii) A directional dependence could be imposed on the selection of stations. Thus if 12 neighbours were to be chosen, a requirement could be made that at least 2 should be drawn from each quadrant. There are situations, however, eg coastal, where this scheme would not work well.

(iv) The neighbours chosen could be weighted in proportion to their correlation with the test station. The weights could be allowed to vary from one month to another.

(v) The length of overlap between stations should be taken into account by including the standard error of the correlation coefficient in the selection criteria. Thus stations could be chosen not according to the correlation coefficient direct, but according to its lower 95% confidence limit, for instance.

Next it is necessary to consider the form of relationship used to estimate the missing observations.

(vi) Should the relation be linear? Strictly, perhaps not, but with the relatively small overlaps commonly available, a linear relationship is the safest assumption to make.

(vii) The calculation of the slope of the linear relation is not entirely straightforward. Let

σ_y = standard deviation of observations from the test station

ox = standard deviation of observations from a neighbour

and r = correlation between the test station and neighbour

The best straight line fit between the stations (the first principal component) will have a slope of oy/ox. A least squared linear regression of y on x, however, will produce a slope of r.oy/ox. This latter will minimise the RMS error of the estimates, but will result in a loss of variance. Thus it is worthwhile setting the slope of the linear relation to both oy/ox and r.oy/ox.

Finally it is necessary to take data for other months into account when making estimates for a particular month. This is achieved by

(viii) Smoothing the monthly values of the weights attached to each neighbour, together with the coefficients of the linear relations with the test station (ie the slope and Δ).

7.2 Formulation of technique

A computer program was written in which the factors discussed above were rationalised as follows:-

(i) The number of neighbours (n) was made to range from 1 to 20.

(ii) The neighbours were selected as follows. For each month, the correlation r with the test station was calculated and converted to Fisher's Z statistic:-

$$Z = \emptyset.5 \left[\ln(1+r) - \ln(1-r) \right]$$

The merit of Z is that its confidence limits are easily calculable from its standard error (E) which is given by

$$E = 1/\sqrt{N-3}$$

where N is the number of pairs of observations. The values of Z associated with each month (Z_m) were averaged over all months to give Z_a and the standard error was calculated from the total number of overlapping observations.

Neighbours were then ranked according to

$$Z' = Z_a - K.E$$

where K was varied from 0 to 8.

Neighbours were not used to make estimates when Z' fell below zero.

(iii) The weights to be attached to the neighbours could be found by solving a set of simultaneous equations which minimised RMS errors. As the variables (neighbours) are highly correlated, however, the equations will be ill-conditioned and the resulting weights unreliable. Four simple schemes, in which the weights W_i were related to the rank i, were tried instead:-

(a) uniform weighting $Wi = 1$

(b) linearly decreasing $Wi = 1 - (i-1)/n$

(c) geometrically decreasing $Wi = 1/i$

(d) exponentially decreasing $Wi = \exp[-(i-1)]$

The weight in a particular month m was then calculated from

$$Wim = Wi . \frac{Zm}{Za}$$

For each month, estimates of missing data were made from linear relations with all the neighbours selected. The final estimate was a weighted average of the estimates from all the neighbours, in which

(iv) The slope of the linear relations were calculated from oy/ox and r.oy/ox.

(v) The monthly values of the weights, slopes, and Δ were smoothed by a j-point filter (involving j months) in which the coefficient of each month was derived from a level of Pascals triangle (eg 1: 4: 6: 4: 1). This gave a close approximation to a Guassian filter, which provides very effective smoothing (see Lee, 1981). The variable j was made to range from 1 (no smoothing) to 11 (approximately equivalent to a running mean of 4 months).

7.3 Optimisation of technique on monthly minima for January

The variables listed above were optimised for monthly min in January using the 10 stations named in section 3. Observations for the stations were supplied

in turn for the years 1959-63, 1975-79, 1959-68, 1970-79, 1959-73 and 1965-79. The results presented below represent averages over the 6 overlapping periods and 10 test stations.

For each of the weighting functions used, table 3 shows how the RMS errors of the estimates vary as the number of neighbours is increased from 1 to 20. In this table, the neighbours were selected by subtracting 4 standard errors from Z, the slope of the linear relations was set to oy/ox, and 7-point smoothing was employed on the slope and Δ . When equal weights were attached to the neighbours, table 3 shows how a substantial reduction in the RMS error (from 1.54°C to 1.27°C) was achieved by increasing the number of neighbours from 1 to 8. As the weighting functions gave less and less weight to the neighbours, more of them needed to be taken into account before the lowest RMS error was attained. The 'geometrically decreasing' weighting function was clearly the best of those tried.

For the geometrically decreasing weighting function, table 4 examines

(i) the method of calculating the slope of the relation between stations, and

(ii) the degree of seasonal smoothing applied to the slope and Δ .

When only one neighbour is used, calculating the slope as r.oy/ox is marginally superior to setting it to oy/ox. When 12 neighbours are used, however, the latter approach is clearly the better. This is especially true for estimating extremes, since r.oy/ox underestimates the variance.

Table 4 shows that the introduction of a modest amount of smoothing to the slope and Δ reduced RMS errors by 0.13°C. Provided there is some smoothing, however, the actual degree of smoothing (within the range applied here) is unimportant. When neighbours were allowed to have different weights in different months, smoothing of these made no difference to the RMS errors. This is understandable, as only small adjustments to the overall weighting function would be made. All the figures presented here were obtained when the weight attached to a neighbour was made the same in all months.

When 23 neighbours are accorded a geometrically decreasing weighting function, the effects of 7 point Gaussian smoothing on the slope and Δ are illustrated in table 5. It is evident that smoothing of Δ is more important than smoothing of the slope, and that smoothing the slope yields only modest improvements over setting it to unity (as in the traditional method). Note that when one neighbour is used and no smoothing is applied to the slope or Δ, the RMS error is 1.68. In the traditional method, in which the slope is set to unity, this is reduced to 1.59.

All the figures presented so far have been obtained when neighbours were selected by subtracting 4 standard errors from Z. The minimum RMS error of 1.21 found so far could be reduced to 1.20 by subtracting 6 standard errors (E) from Z. This may seem to be a large number of standard errors but note that E is based on the total length of overlap available, ie when data from all months are pooled together. If E were calculated from data for only one month of the year, then the number of standard errors subtracted would be much smaller (typically, $\sqrt{12}$ times smaller). If no standard errors were subtracted from Z, the RMS error rises to 1.27.

The above findings may be summarized as follows. When only one neighbour is used, and the slope and Δ are unsmoothed, the RMS error of estimates is 1.68. This can be reduced to 1.20 with contributions, acting independently, as follows:-

(i) Using more neighbours (0.27)

(ii) Smoothing Δ (0.11)

(iii) Smoothing the slope (0.07) [only 0.02 gain over assuming unity]

(iv) Taking into account missing data in the selection of neighbours (0.07)

(v) Weighting the ith neighbour according to 1/i (0.06)

(vi) Calculating the slope from oy/ox (0.04) [but more important in extremes]

(vii) allowing different weightings (0.00)
of neighbours in different months

The optimisation of the technique for monthly min in January may thus be described as

(a) 12 neighbours weighted according to 1/i and selected by subtracting 6 standard errors from Fishers Z transformation of the correlation coefficient.

(b) 7-point Gaussian smoothing of the slope (calculated as oy/ox) and Δ .

Because the neighbouring stations will have incomplete data, the number of neighbours available to form the estimates in any given year will, in general, be less than n. In fact, with 30% missing data (as in this exercise), it will be of the order of 0.7 n. If the number of stations for which $Z^{\prime}$ exceeds zero is less than n, then it will be even less.

It will be evident that the number of neighbours used to form estimates is dependent upon the data array supplied to the routine. If this contained all available data and missing observations amounted to 50%, then the number of neighbours used to form estimates in any given year would be approximately 0.5 n. The results presented in this paper would then apply to this data set if the quoted values of n are multiplied by 0.7/0.5.

7.4 Accuracy of estimates

Estimates of missing data were made for monthly and daily minima in January, monthly and daily maxima in July, and monthly sun in June and December. Although the technique was optimised only for monthly min in January, it has been applied without modification to the other climatic parameters. The estimates were made for the 10 stations described in section 3 and the 6 overlapping periods used in section 7.3. The statistics presented in this section include the term 'mean error'. This is used to describe the arithmetic mean of a set of errors, irrespective of their sign.

The errors of the estimates are expressed as a percentage of the standard deviation of the observations in table 6. The mean errors are approximately 10% of the standard deviation for temperature and 20% for sunshine. Note that there is

only a modest (15%) improvement in the estimates as the length of data increases from 5 years to 15 years.

The mean errors appropriate to 30 year averages are presented in table 7. These were obtained by multiplying the figures given in table 6 by the fraction of data missing in the averaging period. When 15 years of data are available, mean errors are less than 0.1°C for daily max and min, less than 0.2°C for monthly max and min, less than 0.5°C for extreme max and min, and around 3 hours for June sunshine. The mean error is only 80 per cent of the standard error, and so for large sample sizes, 95% confidence limits lie close to 2.5 mean errors. As only small sample sizes were available in this work, 95% confidence limits are best regarded as being around 3 times the mean errors quoted. The figures for extreme max and min were only derived from 21 years of data, and will underestimate the errors associated with 30 year extremes.

Table 7 gives errors averaged over 10 stations for individual months. Errors in other months are likely to vary with the standard deviation, modified by the correlation between neighbours (see figs 3 and 4). Some idea of the differences likely to be experienced from one part of the country to another are provided in table 8. In lowland areas where the stations are relatively close, errors are likely to be only two-thirds of those quoted, while in the interior of Scotland, where the station density is low and the topography complicated, values around 1.4 times as great are to be expected.

8. Comparison of estimates obtained from various techniques

The RMS errors of estimates obtained from the traditional and proposed techniques are compared with those obtained from principal component analysis (Spackman, 1980) and linear interpolation (Hopkins, 1977) in table 9. In making comparisons, the following points should be borne in mind:-

(i) Figures for the traditional, proposed and Spackman techniques refer to Oxford, while those for Hopkins are based on a station network in East Anglia with assumed separations of 20 km for temperature and 25 km for sunshine.

(ii) For Spackman and Hopkins, the entries for monthly max and min refer

to the RMS errors of daily estimates. Values for mean daily max and min and monthly sun have been obtained from the daily values using a relation based on serial correlation. The values used, together with the equation in which they were inserted, are those quoted by Hopkins (1977).

(iii) Hopkins' values were obtained from an average record length of 10 years, with observations being extracted every third day. For temperature, the winter values were drawn from January and February, while the summer values refer to July and August. The results are therefore based on a sample size of 200.

Spackman's values were obtained by extracting observations every third day from 5 years of data (1973-77). For any month, therefore, the sample size is 50.

The traditional and proposed figures were based on extracting one observation per month from an average record length of 10 years. For a given month, this gives a sample size of 10.

(iv) The traditional and proposed methods were both tested on various periods between 1959 and 1979, while the Spackman technique was tested on the years 1977-1980. The Hopkins analysis was based on data in the period 1959-74.

Table 9 shows that at places like Oxford, there is little to chose between the proposed technique and those of Spackman and Hopkins. They all represent considerable improvements on the traditional method.

A more general comparison between the traditional, proposed and Spackman techniques is presented in table 10, where the figures represent errors meaned over the 10 stations named in section 3. Errors associated with the proposed technique are about 72% of those obtained from the traditional method. For temperature, the proposed technique produces better estimates than principal component analysis, but for sunshine the latter technique is clearly superior. This may be ascribed to the poorer station network and greater correlation decay for sunshine than for temperature.

Finally, for monthly minima in January at Corwen, table 11 confirms that

estimates obtained from option SINGLE of program BMDPAM are worse than those obtained from the other methods examined.

9. Conclusions

It is very difficult to write a generalised program to estimate missing data. An obvious starting point is the correlation matrix, but difficulties are created by variations in the length of overlap between variables, leading in extreme cases to missing correlations (obtained when variables fail to overlap). Loss of variance is a general problem associated with minimising RMS errors, and in climatology, the seasonal variation makes it difficult for all the data to be taken into account. The program BMDPAM suffered from all these difficulties.

It is suggested that the best general technique for estimating missing values among a set of highly correlated variables will be based on principal component analysis. The high quality correlation matrix required can probably be obtained from the sample correlations combined with a maximum likelihood routine to estimate the missing correlations. Alternatively, a simple estimating procedure (such as linear regression against the most highly correlated variable) may be used to produce a preliminary set of complete data, and the correlation matrix can be calculated from that.

A statistically simple technique, relying on the availability of a large number of highly correlated neighbours, was devised to produce estimates of climatological data in the UK. It was found to give similar results to an eigenvector scheme used for quality control purposes, so further refinement of the estimates seemed unnecessary. Errors of estimates produced using this technique were about 70% of those obtained using traditional methods. If 10 years of data are available, mean errors of 30 year averages of 0.1°C for daily max and min and 4 hours for June sun are attainable.

As a result, the following criteria are suggested as minimum data requirements for the computation of 30 year averages:

(i) 5 years for daily max and min except in the more data sparse areas, where 10 years are required.

(ii) 10 years for monthly max and min, and monthly sun.

(iii) 15 or 20 years for extreme max and min.

These suggestions are subjective impressions based on the above work and, in practice, should be determined by reference to required levels of accuracy. Quantitative written information on this subject is, however, difficult to find. The most important point is probably that, when an average is computed from incomplete data, some idea of the errors involved will now be known.

Acknowledgement

The BMDP suite of programs were developed at the Health Sciences Computing Facility, UCLA, which was sponsored by NIH Special Research Resources Grant RR3.

References

Dixon, W J 1981 BMDP Statistical Software 1981. Univ Calif Press, Los Angeles.

Dixon, W J & M B Brown 1979 BMDP-79. Biomedical Computer Programs. P series. Univ Calif Press, Los Angeles.

Hopkins, J S 1977 The spatial variability of temperature and sunshine over uniform terrain. Met Mag, 106, pp 278-292.

Jagannathan, P, R Arlery, H Tenkate & M V Zavarina 1967 A note on climatological normals. WMO Tech Note No 84, TP 108, WMO No 208.

Kendall, M G 1975 Multivariate Analysis. Charles Griffin & Co., London.

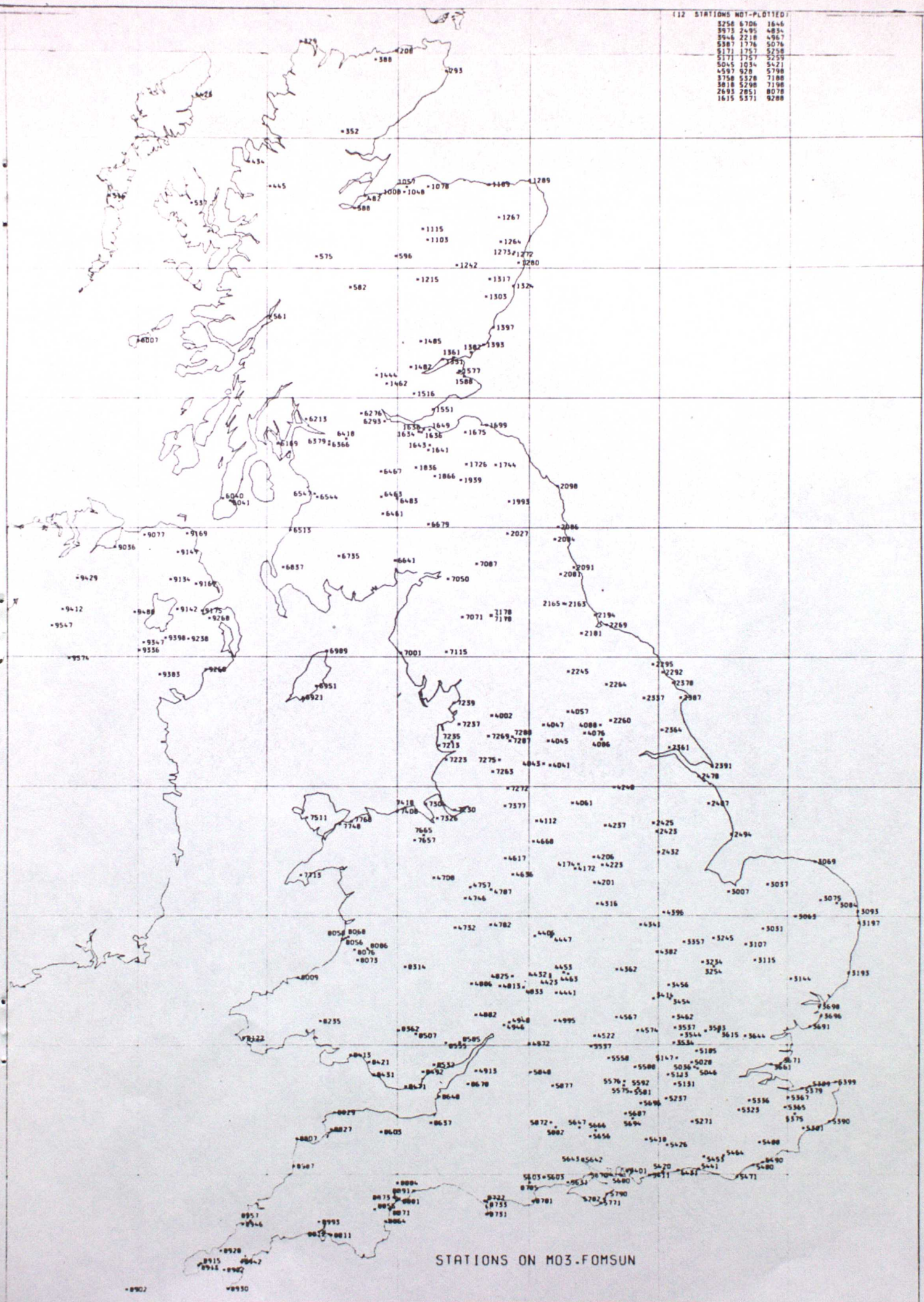
Lee, A C L 1981 Smoothing and filtering of meteorological data. Met Mag, 110, pp 115-132.

Meisner, B N 1979 Ridge regression-time extrapolation applied to Hawaiian rainfall normals. J App Met, 18, pp 904-912.

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

Fig 1 - Distribution of temperature stations

Fig 2 - Distribution of sunshine stations.



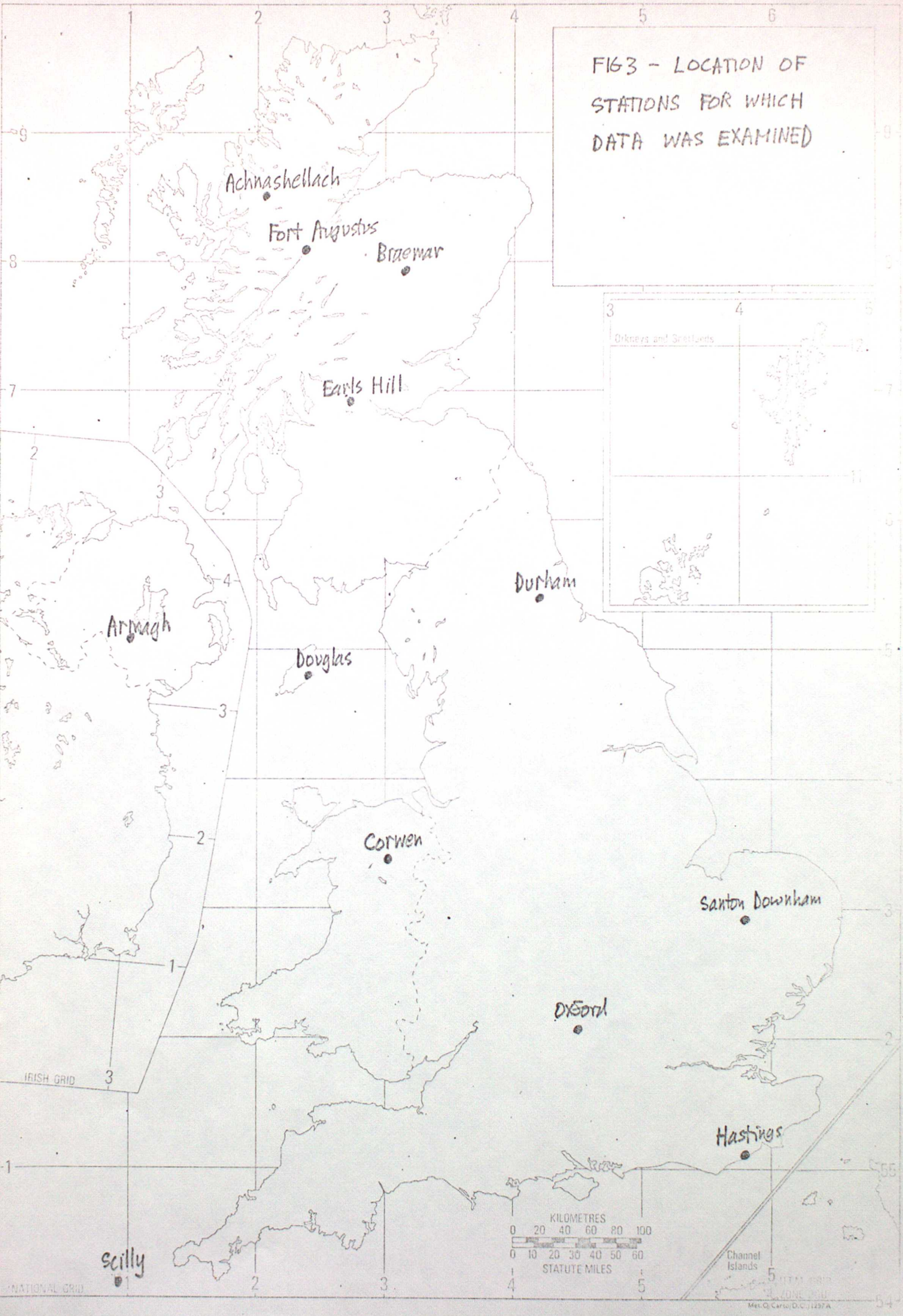


FIG3 - LOCATION OF STATIONS FOR WHICH DATA WAS EXAMINED
Achnashellach
Fort Augustus
Braemar
Earls Hill
Durham
Armagh
Douglas
Corwen
Santon Downham
Oxford
Hastings
Scilly
Orkneys and Shetlands
IRISH GRID
NATIONAL GRID
KILOMETRES
0 20 40 60 80 100
0 10 20 30 40 50 60
STATUTE MILES
Channel Islands
UTM GRID
ZONE 30U

FIG 4 - CORRELATIONS BETWEEN NEIGHBOURING STATIONS

MEAN OF 10 STATIONS IN PERIOD 1959-1979

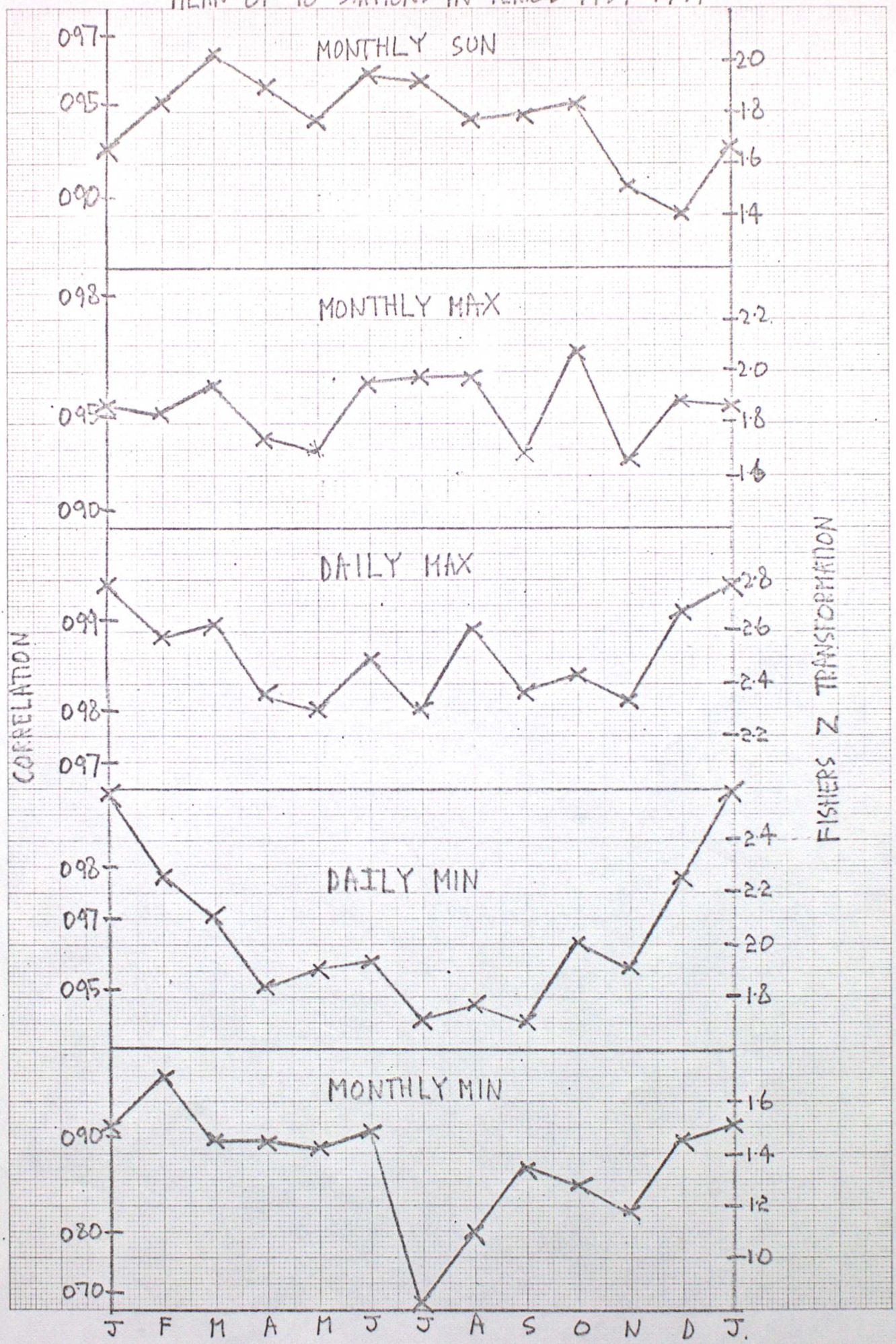


FIG 5 - STANDARD DEVIATION OF SUNSHINE & TEMPERATURE

MEAN OF 10 STATIONS FOR PERIOD 1959-1979

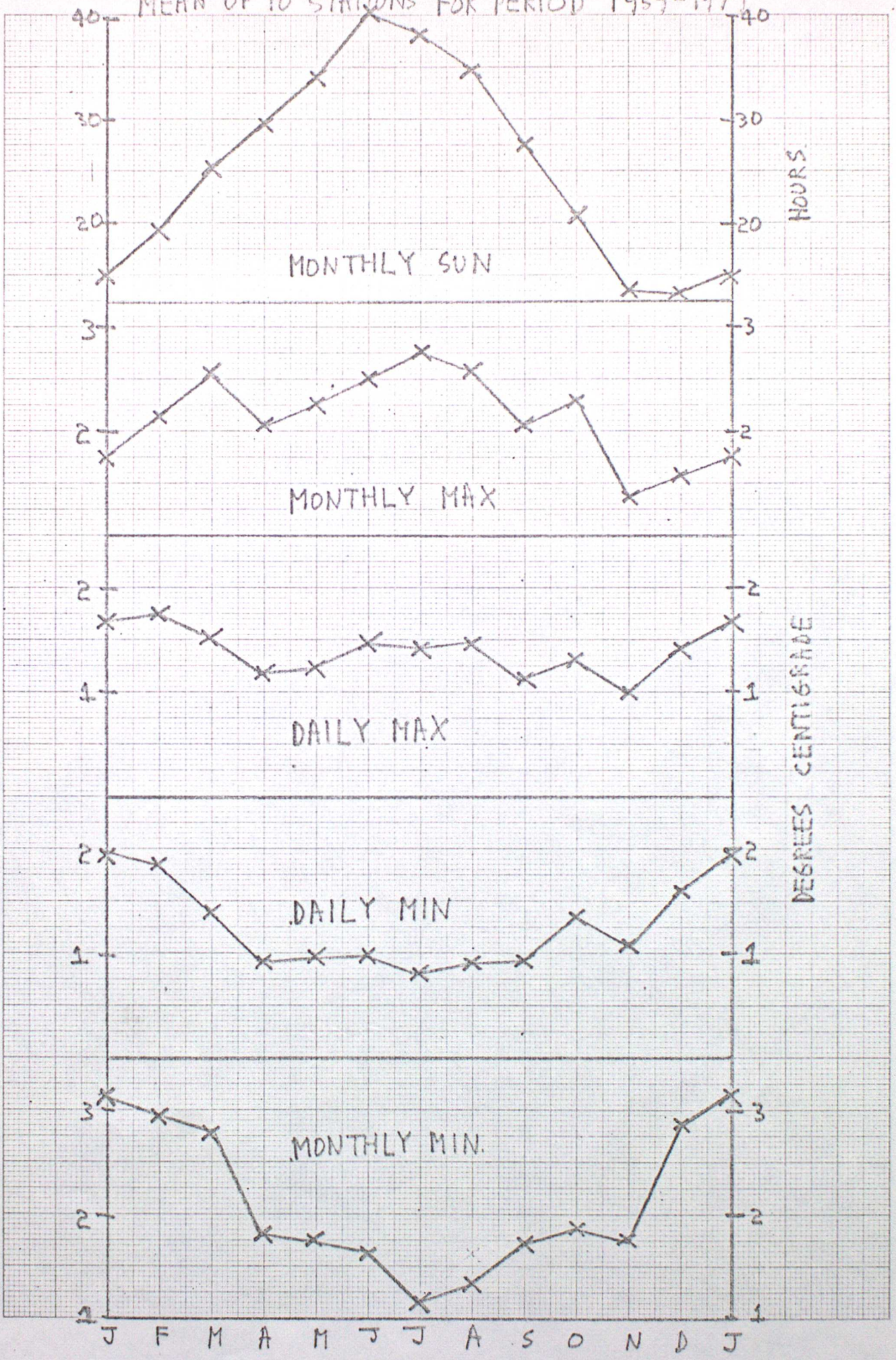


FIG 6 – 'SEASONAL VARIATION OF DIFFERENCES BETWEEN NEIGHBOURING STATIONS.

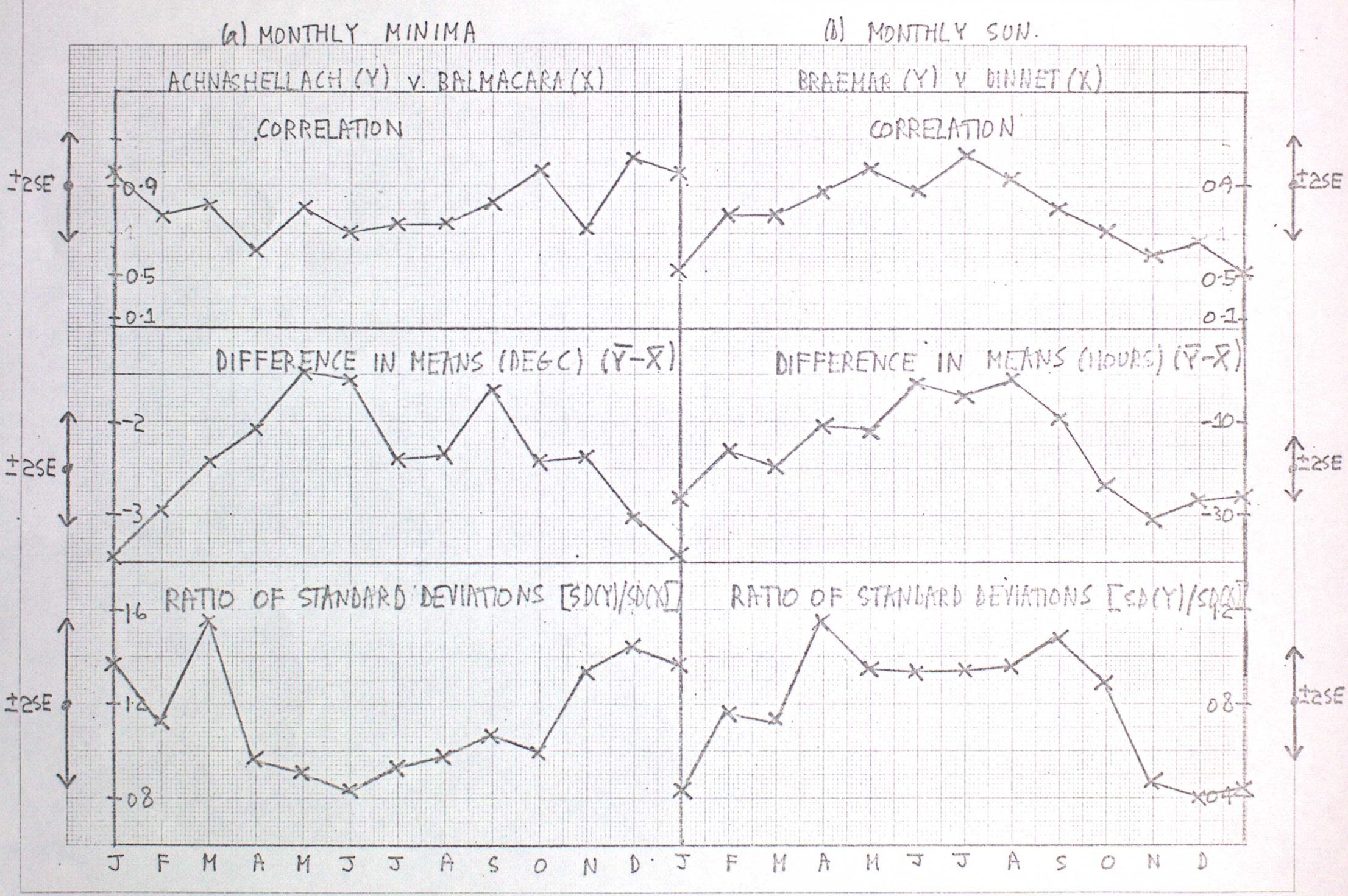


Fig 7 - Estimation of monthly minimum temperatures in January for Corwen by program BMDPAM.

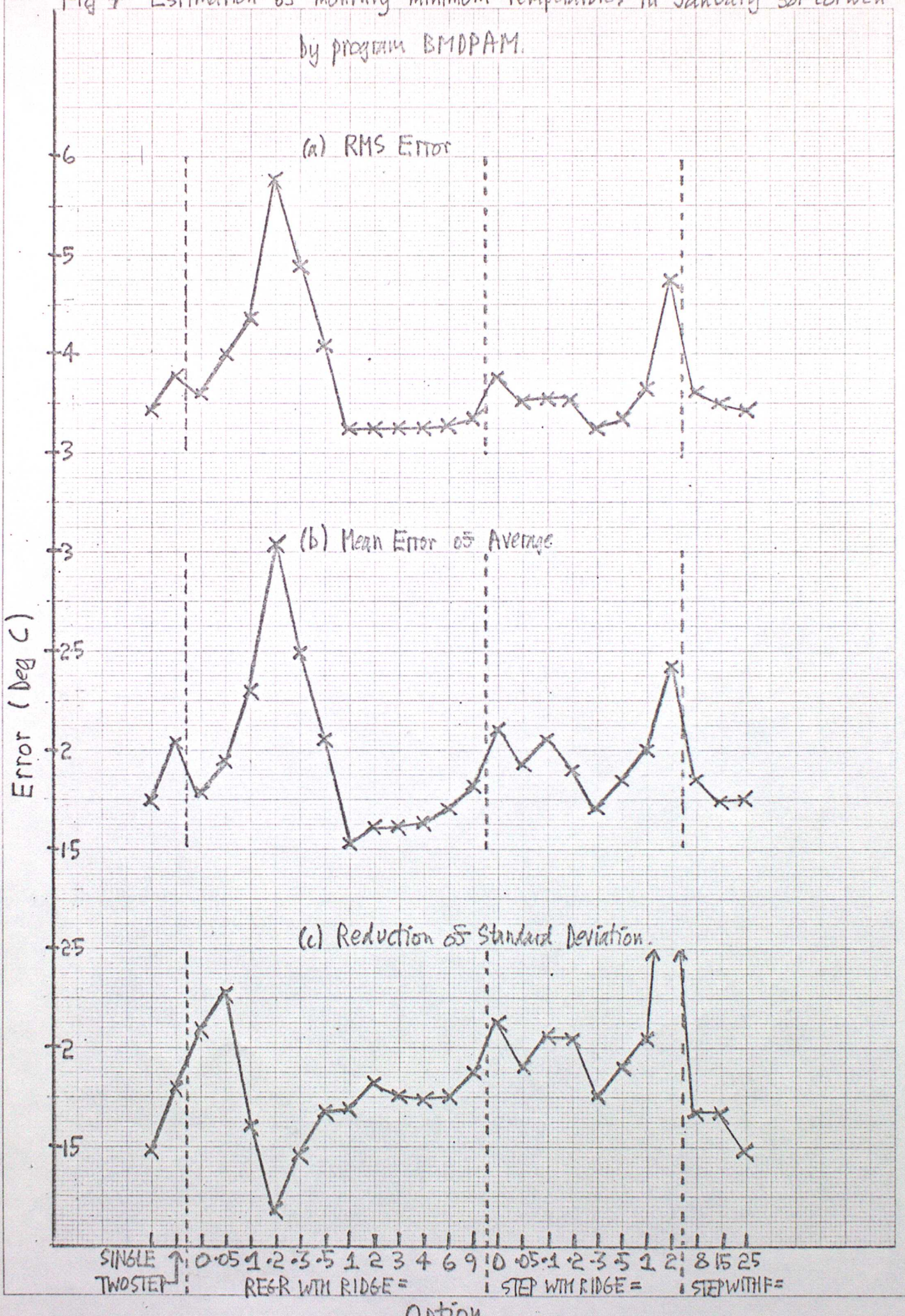


Table 1 - Correlations between neighbouring stations in the climatological network (mean of all months).

Parameter	Average over 10 locations		'Best' location			'Worst' location		
	Highest	6th highest	Name	Highest	6th highest	Name	Highest	6th highest
Monthly Min	0·87	0·81	Oxford	0·90	0·87	Earls Hill	0·80	0·75
Daily Min	0·96	0·94	Oxford	0·98	0·97	Scilly	093	0·91
Daily Max	0·99	097	Oxford	0·995	099	Achnashellach	096	095
Monthly Max	0·95	0·91	Oxford	0·98	0·97	Scilly	079	071
Monthly Sun.	0·94	0·88	Hastings	0·98	095	Fort Augustus	0·89	0·74

Table 2 - Errors of estimates obtained by traditional method

Parameter	RMS Error	Mean Error of Ave.
Extreme min Jan (°C)	–	1·46
Monthly min Jan (°C)	1·59	0·63
Daily min Jan (°C)	0·53	0·28
Daily max Jul (°C)	048	0·23
Monthly max Jul (°C)	1·00	0·40
Extreme max Jul (°C)	–	0·64
Monthly Sun Jun (hrs)	17·4	8·2
Monthly Sun Dec (hrs)	7·8	3·4

Table 3 - RMS errors of estimates of January min obtained by varying the number of neighbours and attaching different weights to the neighbours.
[Slope calculated from $\sigma y / \sigma x$. 7 point smoothing of slope and difference in means; 4 standard errors subtracted from Fishers Z]

Weighting of ith neighbour	Number of neighbours (n).					
	1	3	5	8	12	20
1 (Uniform)	1·54	1·34	1·28	1·27	1·29	1·35
$1-(i-1)/n$ (Linear decrease)	1·54	1·32	1·25	1·23	1·24	1·28
$1/i$ (Geometric decrease)	1·54	1·33	1·25	1·22	1·21	1·22
$\exp[-(i-1)]$ (Exponential decrease)	1·54	1·36	1·31	1·31	1·31	1·31

Table 4 - RMS errors of January min obtained by smoothing over j months.
[Weighting of ith neighbour $= 1/i$; 4 standard errors subtracted from Fishers Z]

	No. of months in smoothing filter.					
	1	3	5	7	9	11
Slope calculated from $\sigma y / \sigma x$						
1 neighbour	1·68	1·57	1·54	1·54	1·54	1·54
12 neighbours	1·34	1·23	1·21	1·21	1·21	1·21
Slope calculated from $r \cdot \sigma y / \sigma x$						
1 neighbour	1·67	1·57	1·54	1·52	1·51	1·51
12 neighbours	1·38	1·28	1·26	1·26	1·25	1·26

Table 5 - Effect of 7-point smoothing on slope and difference in means (Δ).
[12 neighbours; weighting of ith neighbour $= 1/i$; 4 standard errors subtracted from Fishers Z; Slope calculated from $\sigma y / \sigma x$.]

	RMS Error (Deg C)
Δ and slope unsmoothed	1·34
Δ and slope smoothed	1·21
Δ unsmoothed slope smoothed	1·28
Δ smoothed slope unsmoothed	1·23
Δ unsmoothed slope set to unity	1·28
Δ smoothed slope set to unity	1·23

Table 6 - Errors of estimates expressed as a percentage of the standard deviation of the observations

Parameter	RMS Error			Mean Error of Average		
	Length of data					
	15yrs	10yrs	5yrs	15yrs	10yrs	5yrs
Monthly Min Jan	32	40	45	9	10	12
Daily Min Jan	14	19	21	6	10	10
Daily Max July	22	23	25	13	12	13
Monthly Max July	34	31	33	9	9	10
Monthly Sun June	45	36	40	22	17	17
Monthly Sun Dec	53	51	52	21	22	24

Table 7 - Mean Errors of 30 year averages.

Length of data	January			July			Monthly Sun (hrs)	
	Extreme Min (Deg C)	Monthly Min (Deg C)	Daily Min (Deg C)	Daily Max (Deg C)	Monthly Max (Deg C)	Extreme Max (Deg C)	June	Dec
15 years	0·46	0·15	0·08	0·08	0·11	0·27	3·1	1·3
10 years	0·87	0·20	0·11	0·11	0·15	0·37	4·1	1·8
5 years	1·16	0·28	0·14	0·14	0·19	0·47	5·3	2·3

Table 8 - RMS errors of estimates for 10 locations expressed as a percentage of the average for all 10 locations

Location	January		July		Monthly Sun		Location
	Monthly Min	Daily Min	Daily Max	Monthly Max	June	Dec	
Achnashellach	107	132	143	160	107	105	Fort Augustus
Braemar	138	112	170	120	138	92	Braemar
Durham	88	97	100	116	129	90	Durham
Santon Downham	119	106	70	57	100	103	Santon Downham
Oxford	76	82	60	58	67	65	Oxford
Hastings	57	59	70	114	46	50	Hastings
Earls Hill	88	103	100	105	74	185	Earls Hill
Conwen	148	150	110	89	56	69	Douglas
Scilly	91	76	103	117	144	145	Scilly
Armagh	92	91	87	63	138	97	Armagh.

Table 9 - RMS errors of estimates for Oxford

Parameter	Traditional	Proposed	Spackman	Hopkins
Monthly Min Jan (°C)	1·43	0·91	0·74	0·84
Daily Min Jan (°C)	0·40	0·28	0·26	0·30
Daily Max Jul (°C)	0·31	0·18	0·26	0·22
Monthly Max Jul (°C)	0·78	0·44	0·66	0·56
Monthly Sun Jun (hrs)	13·0	9·3	9·0	6·8
Monthly Sun Dec (hrs)	6·8	4·0	4·3	5·1

Table 10 - RMS errors of estimates averaged over 10 stations

Parameter	Traditional	Proposed	Spackman
Monthly Min Jan (°C)	1·59	1·20	1·11
Daily Min Jan (°C)	0·53	0·34	0·39
Daily Max Jul (°C)	0·48	0·30	0·36
Monthly Max Jul (°C)	1·00	0·75	0·90
Monthly Sun Jun (hrs)	17·4	13·9	10·5
Monthly Sun Dec (hrs)	7·8	6·1	4·4

Table 11 - Errors of estimates of monthly min in January for Corwen (deg C)

Method	RMS Error	Mean Error of Average	Reduction of Standard Deviation
BMDPAM option SINGLE	3·43	1·75	1·48
Traditional	2·89	0·99	0·88
Proposed	1·79	0·37	0·41