286

LONDON, METEOROLOGICAL OFFICE.

Met.O.3 Technical Note No.31.

The estimation of missing values from highly correlated data. By CRUMMAY, F.A.

London, Met.Off., Met.O.3 Tech.Note No.31, 1985, 30cm. Pp.30. 4 Refs.

An unofficial document - restriction on first page to be observed.

Met O 3 Technical Note No 31

The estimation of missing values from highly correlated data

by

F. A. Crummay

June 1985

Advisory Services Branch (Met O 3)
Meteorological Office
London Road
Bracknell
Berks RG12 2SZ

Contents

Summary

Generalised routines for the estimation of missing values from highly correlated data are described. The most widely applicable assumes that the data belong to a single population, but an alternative has also been designed for seasonally varying data. The routines were tested on daily minimum temperatures in Northern Ireland and it was found that

i. The fitting of 2 parameters (intercept and slope) rather than one (with the slope assigned to unity) is justified for the majority of stations only when more than 8 months of data are available. For frost hollow stations however, the two parameter fit yields better estimates if more than 5 months of data are available.

ii. Principal component analysis failed to improve on estimates obtained using a constant difference approach, with the notable exception of frost hollow stations.

iii. Nine years of data are required before a single population approach can match the estimates produced by the use of seasonally smoothed regression coefficients

1. Introduction

A notable feature of the last decade has been the rapid expansion of packaged statistical software which has facilitated the implementation of a wide variety of statistical analyses. These analyses generally require a complete set of data as input but this is rarely available and the expansion in statistical software has not been matched by the development of missing data routines to provide the necessary input. A number of simple devices are commonly used to overcome the missing data problem:-

i. replacing the missing observations by a mean value.

ii. ignoring the missing values by using all the available data.

iii. omitting all cases which contain one or more missing values.

None of these options are very satisfactory, however, especially when the data are highly correlated, and the general subject of missing data has received little attention.

The most generalised procedures for the estimation of missing data are likely to be based on the concept of partial correlation and multiple regression with each 'independent' contribution to the variance being estimated from a single variable. The principle of maximum likelihood can be invoked in the derivation of the regression. A general discussion is provided by Beale and Little (1975) while Frane (1978) was responsible for including a missing data routine into the BMDP suite of statistical software. If a large number of highly correlated variables are available an alternative to multiple regression is the formation of a weighted mean of the estimates obtained directly from several of the variables. This forms the basis of a routine developed by Tabony (1983), designed to operate on seasonally ranging climatological data. Tabony also suggested that principal component analysis may have a role to play in the estimation of missing data.

Principal component analysis, described by Kendall (1975), 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 components represent systematic differences between the variables whereas the random differences are usually consigned to higher order components. By reconstituting the data from only the leading components, therefore, the genuine differences between variables are retained while the noise and insignificant effects are ignored. As thus described, the technique seems an ideal means of estimating missing values. It suffers from the disadvantage, however, that a complete and self-consistent correlation matrix is required. This can be obtained from a maximum likelihood approach or by the use of a simple estimating technique to obtain a preliminary set of complete data.

The simple assumption to make regarding the input to a missing value routine is that the data belong to a single population. Many sets of data, however, notably those with a climatological aspect, exhibit a pronounced seasonal variation. This paper describes the development of a routine designed to operate on a single population of data, and which contains a number of options based on simple regression techniques. These are used to provide a preliminary set of complete data for refinement by principal component analysis. The techniques are tested on climatological data and the estimates compared with those obtained from the seasonally smoothing routine of Tabony (1983) in order to provide advice on the relative merits of each.

2. Data

The techniques were tested on daily observations of minimum temperature for 83 stations in Northern Ireland for the ten year period 1974-83. The exclusion from the data set of stations with less than two years observations limited the proportion of missing data to 25%, and the correlation coefficient of the most closely associated station-pairs averaged 0.92. The single population routines were provided only with data for January, giving an array size of 83 stations (variables) x 310 days (cases).

The performance of the techniques was assessed by withholding data from specified stations for a prescribed number of days, making estimates for those days, and then comparing the estimates with the true observations. The accuracy of the technique will vary with the test station - observations from unusual sites such as frost hollows will be more difficult to estimate than those from the more commonly occurring sites. For this reason, the techniques were assessed over seven stations whose site characteristics are described in Table 1. The standard of the estimates also varies considerably from year to year, with cold Januaries such as 1982 presenting considerably more problems than mild months such as occurred in 1975. This is illustrated in Table 2 where the RMS errors using one of the techniques are given when temperatures in each January were estimated from the remaining nine. The years withheld for testing the technique were selected as 1974, 1978 and 1983, all reasonably typical years. Each year was withheld in turn and estimated using one to nine years of the remaining January data.

Table 1 Test Stations in Northern Ireland

Station Number	Station Name	Site Characteristics
9134	Lisnafillan	Frost Hollow
9157	Parkmore Forest	Upland
9169	Ballypatrick Forest	Forest clearing
9288	Helen's Bay	Coastal
9336	Armagh	Small town, lowland site
9375	Tandragee	Semi-urban lowland
9506	Knockaveran	Near lough

Table 2 Mean R.M.S. errors for each January (1974-83) for Ballypatrick Forest, estimates made using 1-9 neighbours in the final estimation process. CLIMREG option used.

No. of neighbours	1	2	3	4	5	6	7	8	9
R.M.S. error (deg C)	1.37	1.25	1.16	1.11	1.11	1.10	1.10	1.10	1.10

3. The single population technique

The main steps in the single population technique are as follows:-

i. An array of data (variables x cases) is supplied and each variable takes a turn at being the predictand.

ii. The correlations between the predictand and the remaining variables are calculated from only those cases for which observations from both the predictand and the variable are available.

iii. The 'best' variables are selected according to the lower 95% confidence limit of their correlation with the predictand. This procedure enables the number of observations available for the computation of the correlation to be taken into account.

iv. Data for the predictand are regressed against those for each of the selected variables in turn and a set of preliminary estimates are obtained.

v. A final estimate is obtained by ataching a weight 1/i to the estimate associated with the i th ranking variable (Tabony, 1983).

The estimates derived from the individual variables were obtained in four ways:-

I. Conventional linear regression. This involves fitting an intercept and a slope of r $\sigma y/\sigma x$ where r is the correlation between the variable and predictand and σx and σy are the standard deviations of the variable and predictand respectively.

II. Linear regression with a slope of $\sigma y/\sigma x$ ie the moderating effect of the correlation is omitted. This is likely to be beneficial in many climatological applications when the variables are merely different measurements of the same element at neighbouring stations. For the case of $\sigma y = \sigma x$, for example, there will be a random scatter about a 1:1 relationship, and the slope of the regression should be 1 rather than r.

III. Linear regression forced through the origin. This is equivalent to setting the slope equal to $\bar{y}/\bar{x}$, where $\bar{x}$ and $\bar{y}$ represent the means of the data samples for the variable and predictand respectively. In climatology, this may be appropriate to rainfall and sunshine data.

IV. Linear regression with the slope set to unity. This corresponds to the imposition of a constant difference $(\bar{y} - \bar{x})$ between the variable and predictand and which, in climatology may be appropriate to temperature.

The single population routine is documented in Appendix 1.

4. The seasonal technique

A routine designed to estimate seasonally ranging climatological data (Tabony, 1983) is documented in Appendix 2. It is very similar to the single population technique described above but embodies the following differences.

i. Only one option is available and estimates are based on linear regression in which the effect of the correlation on the slope is ignored.

ii. The regression coefficients for each variable are calculated separately for each calendar month, and then smoothed over all

iii. The number of variables that can be used is fixed at 12, but the number that is actually used is limited to the subset of variables (from the 12) which are available for the case in question. This makes the number of variables used sensitive to the proportion of missing data. In the single population technique this disadvantage is overcome by searching the data (for a given case) until the number of variables used equals the number specified.

5. Comparison of techniques

There are two main queries relating to the application of the techniques as described above:-

i. Given that the data belong to a single population, which of the four single population options will be the most appropriate to use?

The answer will depend on the amount of data available. In option I the minimum assumptions are made about the data and therefore systematic errors are likely to be smallest. The size of the random error depends on the accuracy with which $\bar{y}$, $\bar{x}$, σy, σx and r can be evaluated. The technique is likely to perform well when a large number of cases are available. In option II the correlation r is assigned a value of 1 (appropriate to climatological applications) but the need to calculate the standard deviations σx and σy maintain the requirement for a large number of cases.

In options III and IV prior information is used to assign a value ($\bar{y}/\bar{x}$ or 1) to the slope and this obviates the need to compute σx and σy. The assignment of a value to the slope invokes a penalty in terms of the systematic error involved, but the need to compute only $\bar{x}$ and $\bar{y}$ has a beneficial effect on the random errors. The technique is therefore likely to perform well when only a small number of cases are available.

ii. Given that the data vary seasonally, what are the relative merits of the seasonally smoothing routine and the separate treatment of each month as a single population?

The smoothing operation will introduce a systematic error but the use of up to twelve times as much data will reduce the random error. The single population approach may therefore prove the best technique if a large amount of data is available, but otherwise the seasonally smoothing technique is likely to be superior.

The techniques compared in this section are, for the single population routine

a. Conventional linear regression (Option I) LINREG

b. Linear regression with slope set to $\sigma y/\sigma x$ (Option II) CLIMREG

c. Linear regression with the slope set to unity = constant difference (Option IV)) CONDIF

In addition estimates were made using

d. Linear regression with the slope set to unity and the use of only one variable. This corresponds to the method traditionally used by the Met. Office in the hand calculation of long period averages of temperature TRAD

e. Seasonally smoothed linear regression SEASON

The number of variables (neighbouring stations) to be used for a single population options was set at 8. This is consistent with a maximum of 12 in the seasonally smoothing routine, which was found to be optimal when only 70% of data were present. It is further justified in Table 3, which displays the RMS errors obtained when from one to nine neighbours were used to estimate observations at Ballypatrick Forest using the CLIMREG option.

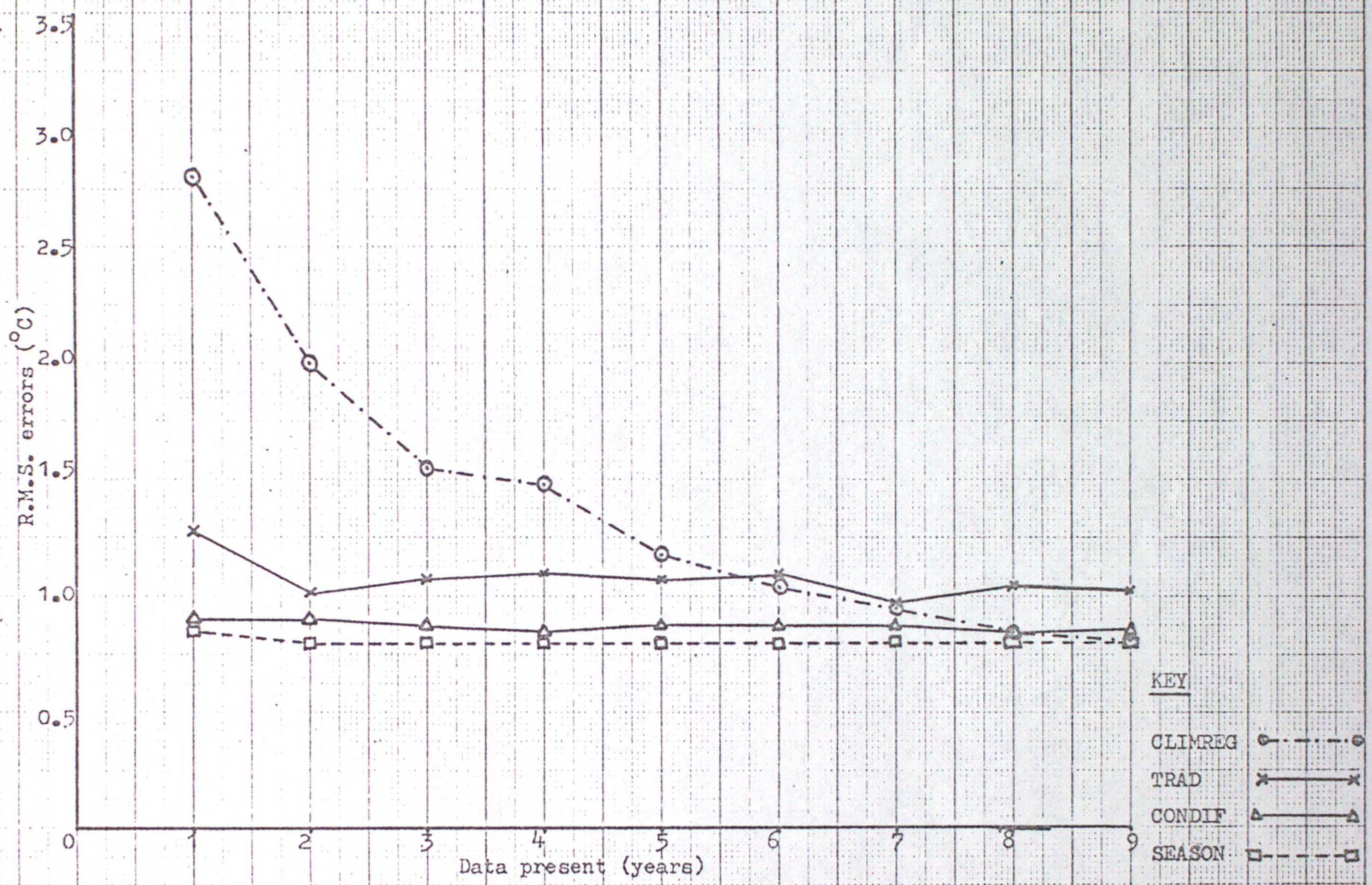
Table 3 The effect of the character of the weather on R.M.S. errors for January. CLIMREG option and 9 months of January data used to estimate each January's data.

Year estimated	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983
R.M.S. error (deg C)	1.16	0.80	0.93	1.45	1.22	1.41	1.54	0.85	2.06	1.11

The RMS errors obtained from the various techniques, using from 1 to 9 Januaries to estimate for the 7 test stations, are shown in Fig. 1. The most outstanding feature is the poor results obtained from the single population options which require the evaluation of both slope and intercept - CLIMREG and LINREG (not illustrated). When data from only a single January is available, the need to calculate the standard deviation from only ten independent data points leads to RMS errors of almost 3.0 deg C. As the length of data available increases, the RMS errors decrease steadily to 1 deg C. For the special case of climatological data, the omission of r from the evaluation of the slope leads to a modest (0.1 deg C) reduction in error.

In contrast, the options which set the slope to unity produce errors which are relatively insensitive to the amount of data supplied with the use of eight neighbours (CONDIF) providing an improvement of around 0.2 deg C over the use of only one (TRAD). When determining the regression slope from the data, approximately eight

Figure 1 R.M.S. errors for 7 test stations and 3 test Januaries (1974,1978,1983) using 1-9 years of data to estimate missing values.



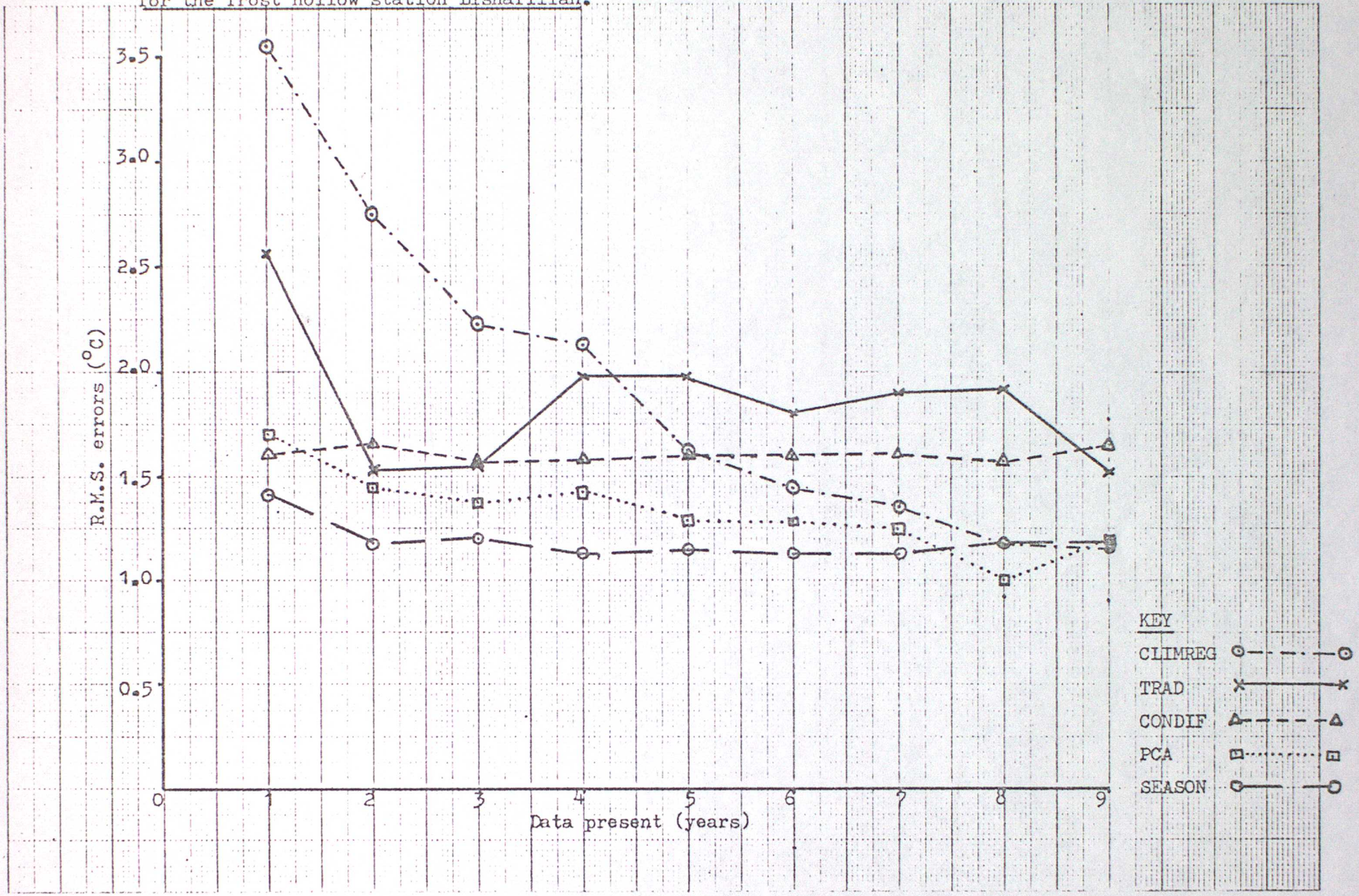
years of data (240 days, of which about 80 are independent) are required to produce estimates which are as accurate as the constant difference approach.

Marginally the best estimates (RMS ~ 0.8 deg C) are obtained by using data from all seasons and smoothing the regression coefficients over all months (SEASON). The benefits are again greatest when only a small quantity of data is available and approximately nine years of data (90 independent days) are required before the single population technique yields a similar accuracy.

The assignment of the slope to unity produces good results since, for the majority of highly correlated climatological station pairs, it is a good assumption to make. For unusual sites such as frost hollows however, the slope is likely to differ from unity and the balance of advantage shifts towards the full regression option. This is illustrated in Fig. 2 where the RMS errors associated with the frost hollow station (Lisnafillan) are examined. Not surprisingly, the errors are larger than for the majority of stations, and the CLIMREG option now provides better estimates than CONDIF when more than 5 years of data are available.

With frost hollow sites, the advantage of using several neighbours over just one might be expected to be less than with the majority of stations. The use of only the best neighbour (possibly another frost-hollow) might be better than the use of several less highly correlated stations. These remarks apply particularly to the

Figure 2 R.M.S. errors for 3 test Januaries (1974, 1978, 1983), estimated using 1-9 years of January data for the frost hollow station Lisnafillan.



constant difference approach, which is less able than the full regression to cope with station-pairs of differing site characteristics. It is therefore interesting to note that for Lisnafillan, Fig. 2 shows that the constant difference approach with the use of eight neighbours (CONDIF) has the advantage over the use of only one, as in the traditional method.

6. Computing Costs

The computing costs depend mainly upon.

i. The total number of variables supplied, and to a lesser extent the number of cases.

ii. The number of variables for which estimates are required.

An idea of the costs involved is given in Table 4 for an array size of 83 variables x 310 cases for the single population options, and twelve times as much for the seasonal techniques. The number of cases estimated was fixed at 10% of those supplied, but this is not a factor of great importance. The various options of the single population technique all incur similar costs.

Table 4 Costs incurred in the use of the estimation techniques, 10 years of data is supplied of which one is replaced and estimated.

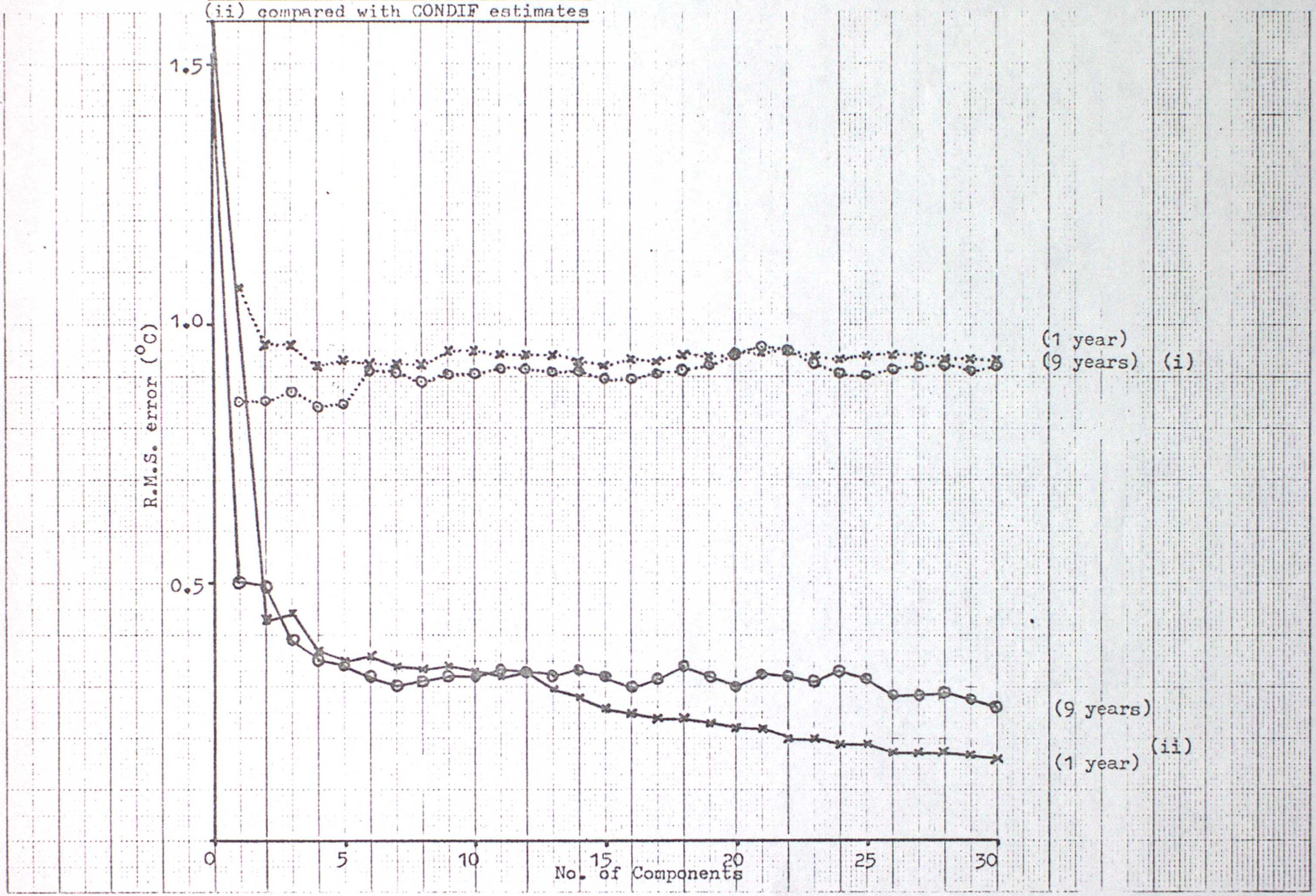
Method	No. of Stations	Period to be estimated	Cost (Units)
Single pop'n options	1	1 month	5.7
"	7	"	5.7
"	83	"	15.88
"	1	12 months	31.94
"	7	"	35.38
"	83	"	approx 17.00
SEASON	1	1 month or 12 months	9.46
"	7	"	23.25
"	83	"	approx 145.00

If estimates for only a single calendar month are required then the single population technique is much cheaper than the seasonal. For the former routine costs ranged from 6 to 16 units as the number of variables to be estimated increased from 1 to 83, while for the seasonal technique the corresponding range is from 9 to 145. If estimates for all calendar months are required, however, then the seasonal technique becomes the cheapest. The costs of that technique are the same as for a single calendar month, while those for the single population approach have increased to between 32 and 170 units. For the seasonal technique, however, storage requirements may be excessive.

7. Principal Component Analysis

The constant difference option (CONDIF) was used to provide a preliminary set of complete data which were re-estimated from the leading eigenvectors of a principal component analysis. A major problem in the application of principal component analysis is the choice of the number of components to use. In general, this will increase with the number of variables and decrease with the degree of association between them; the number of cases will also be a factor if they are less than the number of variables. The effect of varying the number of components on the accuracy of the estimates of minimum temperature in Northern Ireland is illustrated in Fig 3. If the estimates are compared with the data supplied to the principal component routine, the differences decrease steadily to zero as all (83) components are used. If the estimates are compared with the withheld values, a decrease in RMS error is associated with the first few components. As further components are added a plateau is apparent in which the quality of the estimates is insensitive to the number of components used. Although the limited nature of this study does not permit much detailed advice on the choice of the best number of components to use, the lack of sensitivity displayed in Fig. 3 is encouraging in that it indicates that the number chosen is unlikely to be critical. A choice of four components was made for the present work.

Figure 3 Mean R.M.S. errors for 7 test stations based on P.C.A. using 1 or 9 years of data. 2 options present
(i) compared with the genuine data
(ii) compared with CONDIF estimates

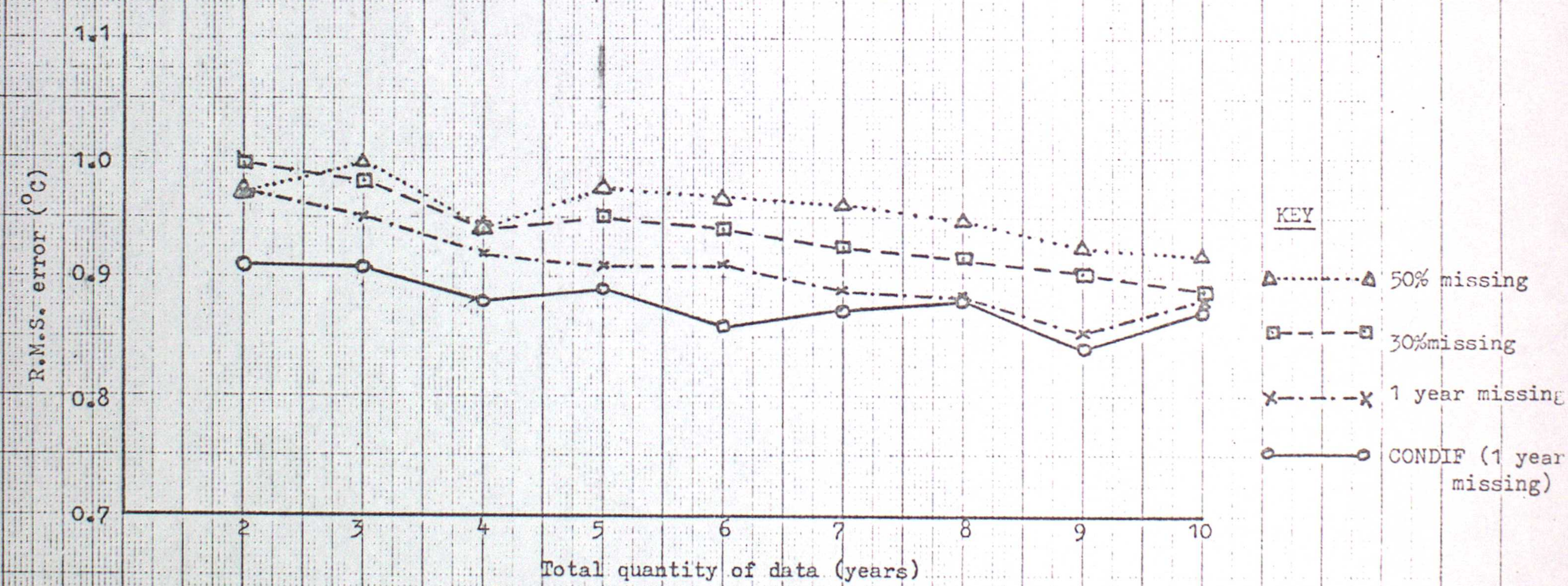


The ability of the principal components technique to provide good estimates will clearly be influenced by the amount of data which has been prehestimated, both in the data array as a whole and in particular for the test station itself. The effect is however, shown to be modest in Fig. 4, where the gradual decrease in RMS error with the number of Januaries data supplied is displayed with the amount of pre-estimated data for the test station varying from 50% to 30% to one year.

Fig. 4 also displays the RMS errors from the constant difference technique, from which it becomes apparent that the principal component technique has failed to improve on the preliminary estimates. There is, however, an important qualification - the quality of the estimates for the constant difference technique vary considerably from station to station, being much worse for the frost hollow site than the other stations. For the principal component technique, the variation in the standard of the estimates between stations is much less, and the values produced for the frost hollow station are superior to those provided by the constant difference approach (Fig.2). RMS errors for Lisnafillan obtained when data from nine Januaries were available are 1.65 deg C for the constant difference approach compared to 1.20 deg C for both the principal component and CLIMREG techniques.

It is possible that a principal component analysis based on a correlation matrix derived from a maximum likelihood procedure (as opposed to one based on pre-estimated data) would perform better. There does not seem to be much scope for improvement, however, over the estimates obtained when only 10% (1 year) of the test station's observations have been pre-estimated.

Figure 4. Mean RMS errors for CONDIF and PCA estimates using different percentages of missing data



8. Conclusions

Generalised routines have been written for the estimation of missing values from highly correlated data. The most widely applicable regards the data as belonging to a single population and contains a number of options, while another is designed for seasonally varying data. Application of the routines to minimum temperatures in Northern Ireland shows that

i. The use of 8 neighbours in the constant difference option led to a reduction in RMS error of about 0.2 deg C over the use of only one neighbour, as in the traditional approach. This statement remains true for the frost hollow station examined (when its veracity might have been questioned).

ii. For the majority of stations, the fitting of two parameters as in conventional linear regression requires about 8 months of daily data before it can improve on the constant difference approximation in which the slope is set to unity. For a typical frost hollow site however, for which this latter approximation is less valid, a two parameter fit yields better estimates when more than 5 months daily data are supplied.

iii. More generally, it may be said that about 80 independent data points are required before a reliable fit can be obtained from a conventional linear regression. If less than 30 independent observations are available, considerable improvements can be made if any prior information is available which will lead to the need for the data sample to estimate only one parameter.

iv. Principal components analysis failed to provide an overall improvement in the quality of the estimates provided by the constant difference approach. A notable exception occurred for the frost hollow site, where improvements were obtained on the 'pre-estimates' provided by the constant difference technique.

v. For seasonally varying data, reliable fits can be obtained through the use of seasonally smoothed regression coefficients. More than nine years of data are required before the accuracy achieved by this technique can be matched by the single population approach.

References

Beale E. M. L. and Little R. J. A. 1975 'Missing values in multivariate analysis'. J. Roy. Statist. Soc., B.37, 129-145.

Frane J. W. 1978 'Missing data and BMDP: some pragmatic approaches'. BMDP-77 Tech Report No 45, Dept of Biomathematics, UCLA.

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

Tabony R. C. 1983 'The estimation of missing climatological data', J of Climatology. Vol 3, pp 297-314.

Appendix 1 The estimation of missing data belong to a single population

Subroutine MISDAT

Source MO3.BRANCH.FORT(QDMISDAT)

Object MO3.BRANCH.OBJ(QDMISDAT)

Purpose To estimate missing values in an array (cases, variables) ie (day, station) of highly correlated (eg climatological) data.

General Descriptions

i. Estimates are made for the first ITEST variables in an array (cases, variables) of data in which missing observations are represented by any values less than THRESH.

ii. Each of the first ITEST variables takes a turn at being the predictand. Any variables which contain less than IPAIR paired (overlapping) observations with the predictand are rejected. The remaining variables are ranked (according to the lower 95% confidence limit of Fishers Z statistic), and the top IPRELN (recommended value, 20) are retained to perform the initial estimation.

iii. Data for the predictand are regressed against that for each of the selected variables - the regressions may be calculated using the following options.

1. Conventional linear regression. Slope = r $\sigma y/\sigma x$. r = correlation between the variable and predictand, σx and σy are the standard deviation of the variable and predictand.

2. Linear regression with the slope set to $\sigma y/\sigma x$.

3. Linear regression forced through the origin. Suitable for rainfall or sunshine data.

4. Linear regression with the slope set to unity. Suitable for temperature.

Only one option may be selected (IOPT) with the estimates replacing the missing values in the dataset on exit.

iv. The best IFINAL variables (recommended value 8) with data present for a given case (day) are chosen to form the final estimates. These are obtained from a weighted average of the initial estimates a weight 1/i being attached to the value associated with the i th ranked neighbour.

Method of Use

CALL MISDAT (IVAR, ICASE, IDCNN, IPAIR, IOPT, ITEST, IFINAL, IPRELN, DAT, THRESH, JDCNN)

Supplied by subroutine

IVAR =	Integer *4 =	Number of variables (stations). Missing data will only be eliminated from the first ITEST variables.
ICASE =	Integer *4 =	Number of cases (days)
IDCNN =	Integer *4 =	Array of dimension (IVAR) containing variable identifier (station number)
IPAIR =	Integer *4 =	Minimum number of paired (overlapping) observations permitted.
IOPT =	Integer *4 =	Regression option required (in range 1-4 above).
ITEST =	Integer *4 =	Number of variables (stations) for which estimates will be made. These must be the first ITEST variables supplied.
IFINAL =	Integer *4 =	Number of variables to act as predictors (neighbours) for the final estimation process. (Recommended value for climatology = 8).
IPRELN =	Integer *4 =	Number of most highly-ranked variables which are to be searched in an attempt to find IFINAL of them present for a given case (day). (Recommended value for climatology = 20).

DAT = Real * 4 = Array of dimension (ICASE, IVAR) containing data in form (day, station). Missing values in first ITEST variables estimated and replaced on exit.

THRESH = Real * 4 = Threshold value below which any values of DAT are deemed to be missing.

Returned by subroutine

JDCNN = Integer *4 = Array of dimension (ITEST, IPRELN) containing the variable identifiers (station numbers) of the IPRELN most highly ranked neighbours, for each of the ITEST variables.

DAT = Real *4 = Array of dimension (ICASE, IVAR) containing data in form (day, station). Estimated values indicated by the addition of 10,000.

JCL

```
   :
   :
   :
//LKED.FAD  DD  DSN= M03.BRANCH.OBJ,DISP=SHR
//LKED.SYSIN DD *
       INCLUDE  FAD(QDMISDAT)
   :
   :
   :
```

Appendix 2 - The estimation of missing values in seasonally varying data

Soubroutine MISIN

Source M03.BRANCH.FORT(QQMISIN)

Object M03.BRANCH.OBJ(QQMISIN)

Purpose To estimate missing values in an array (case, month, variable) (eg year, month, station) of climatological data.

General Description

(i) Estimates are made for the first INSTN variables in an array (cases, month, variables) of data in which missing observations are represented by -32768.

(ii) Each of the first INSTN variables acts as the predictand, and variables which contain less than four years data will not have the missing values estimated and a message to this effect is printed.

(iii) Correlations between the predictand and the remaining variables are calculated for each month, converted to Fishers Z statistic, and meaned over all months. The variables are then ranked (according to the lower 95% confidence limit of Fishers Z statistic) and the 'best' 12 variables (neighbours) selected.

(iv) For each of these variables and for each month, data from the predictand are regressed against that form the neighbours. (The gradient is set equal to the ratio of the standard deviations).

(v) The regression coefficients are smoothed over all months (using a 7 point filter) and the smoothed values used to estimate any missing values. A weight of 1/i is attached to each ith ranked variable and the final estimate obtained from the weighted mean of those values obtained from the 12 'best' neighbours.

Method of Use

```
CALL MISIN (IDCNN,INR,INSTN,NDAT,MM,INYR,PDAT,JDCNN)
```

Supplied to subroutine

IDCNN = Integer *4 = Array of dimensions INR containing DCNN numbers of stations for which data are supplied.

INR = Integer *4 = Total number of stations supplied. Missing data will only be eliminated from the first INSTN stations. The remainder act only as neighbours.

INSTN = Integer *4 = Number of stations for which estimates are to be made.

NDAT = Integer *2 = Array of dimensions (INYR,MM,INR) containing data in form (year, month, station).

MM = Integer *4 = 12

INYR = Integer *4 = Number of years data supplied.

Returned by subroutine

PDAT = Integer *2 = Array of dimensions (INYR,MM,INSTN) containing data from which values have been eliminated.

JDCNN = Integer *4 = Array of dimensions (MM,INSTN) containing DCNN numbers of stations used as neighbours. The array is arranged as [neighbours (rank 1 to 12), station supplied (i = 1 to INSTN)].

JCL

```
  :
  .
  :
//LKED.RCT DD DSN= M03.BRANCH.OBJ,DISP=SHR
//LKED.SYSIN DD *
        INCLUDE RCT(QQMISIN,QQMESK)
          :
          .
          :
```