TECHNICAL MEMORANDUM

63

A screening multiple discriminant analysis program

L.J. Wilson

Operations Department

1982

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen

CONTENTS

1. INTRODUCTION

Multiple Discriminant Analysis (MDA) is a linear statistical procedure to provide an objective basis for categorising a predictand into one of several groups according to given predictor values. In meteorology the method is ideally suited to those applications where the predictand meteorological variable is characteristically non-continuous, or is usually categorised for operational reasons. Examples of meteorological elements which are suited to prediction by discriminant analysis are precipitation type and amount, cloud amount, ceiling and visibility. MDA may be applied to these elements using the perfect prog or Model Output Statistics (MOS) formulation, or a combination of these, depending on the predictors that are offered.

MDA may be applied to meteorological problems in a manner similar to regression or any other statistical forecasting procedure. Equations are developed using a dependent sample which should be as large as possible, then tested on an independent sample to estimate the forecast skill obtainable. As there is very much overlap in the information content of meteorological variables, predictor screening is advisable to choose the best subset of a large number of available predictors. Screening methods employed in MDA are the same as those used in regression although the statistic used to test for acceptance or rejection is different. The program described below uses the Forward Selection screening method (Draper and Smith, 1981).

Two separate programs are available at ECMWF for the use of MDA: a program called FORWMDA and a program called MDARUN. The former carries out a screening MDA and produces diagnostic output based on the dependent dataset while the latter is designed to produce and verify probability forecasts using the output of the development program. MDARUN can be used on either dependent or independent data samples providing the input data is supplied in the required formats. Details of the required input data formats, and procedures to access the programs are contained in a user guide available from ECMWF. This report describes the development program FORWMDA and demonstrates the use of its output by means of an example.

2. GENERAL DESCRIPTION OF PROCEDURE

The final output of a Multiple Discriminant Analysis is the probability of occurrence of each predictand category. The procedure may be divided into the following main steps:

1. Read and prepare input data for analysis

2. Screen for predictors

3. Calculate discriminant functions for the dependent data

4. Compute all coefficients required for application to independent data

5. Determine probabilities (application)

The analysis is complete after the fourth step, but the program continues with the fifth step using the dependent dataset. The method was first described by Fisher, 1936, and the meteorological application and screening procedure are well documented in Miller, 1962. The program FORWMDA closely follows the parametric procedure described in the Miller publication. The non-parametric procedure requires fewer assumptions, but necessitates re-analysis of the dependent dataset each time a forecast is wanted and is therefore impractical for operational meteorological applications.

2.1 Read and prepare input data for analysis

Unlike regression, the MDA procedure does not explicitly use the predictand in the formulation of the equations. The predictand values are required only to partition the data into categories. The input dataset must be physically reorganised into categories, and it is therefore necessary to store the original dataset in memory, while the database is rearranged. The procedures for data input are described in detail in the user guide.

2.2 Screening for predictors

The screening procedure used in the program follows that suggested by Miller, 1962, with the exception that no explicit statistical test is carried out to discontinue selection. Instead, selection is discontinued when the addition of the best new predictor increases the value of the test statistic by less than 10%.

The test statistic is the Mahalanobis D^2 value, a multivariate generalisation of the student's T-statistic which measures the discriminating power of the predictors in terms of separation of category means. Use of a percentage increase in the test statistic as a cutoff criterion ensures that selection will always occur, but does not guarantee a statistically significant equation. It is therefore worthwhile to do a formal test on the final equation that results.

2.3 Calculation of discriminant functions

Discriminant functions are linear combinations of the predictors for which the separation of the group means is maximised and the within-group dispersion is minimised. They are obtained through an eigenvalue analysis of the within- and between-group dispersion matrices formed from the dependent data. The number of discriminant functions obtained will be one fewer than the number of groups, and they are independent but not orthogonal. They are ordered so that the first contains most of the discriminant power; in fact, the eigenvalues can be used directly in significance tests of the discriminant power contained in the predictors. The eigenvectors obtained from the analysis are the coefficients of the predictors in the discriminant functions.

2.4 Calculation of coefficients required for application to independent data

The parametric procedure requires that certain parameters derived from the dependent data be stored for later application to independent data. These parameters are:

a. The a priori frequencies of occurrence of the groups. These are estimated simply by using the frequencies of occurrence in the dependent sample.

b. The coefficients of the predictors (the discriminant functions).

c. The group means of the discriminant function values in the dependent sample. To obtain these, all the dependent data is transformed to discriminant function values ("discriminant space") by applying the discriminant function equations to each event. Then means are calculated over the dependent data for each category and each discriminant function.

d. The inverse of the within-groups dispersion matrix for the dependent data in discriminant space. The transformed dataset is subjected to the same dispersion computation that was used for the original dataset. The resulting matrix is diagonal to within roundoff error because the discriminant functions are independent.

2.5 Calculation of probabilities

The discriminant functions are used to estimate probabilities of occurrence of each predicted category given a set of predictor values. To make this possible, two assumptions are made:

a. The dependent data is normally distributed within groups.

b. The within-group dispersions are equal among all groups.

If these assumptions are made, the multivariate normal distribution may be used in the computation of probabilities. The parameters of the distribution are the standard variance matrix, and the vector of group means, estimated from the dependent data using the within-groups dispersion matrix in discriminant space and the discriminant function means, respectively. The probabilities are calculated using Bayes' rule, with a priori probabilities also estimated from the dependent dataset (see Appendix A).

3. DESCRIPTION OF THE PROGRAM

3.1 Input

a. Dependent dataset - the program expects error-free dependent dataset of, at present, up to 200 events, consisting of one predictand and up to 99 predictors. This array is prefaced with two records, the first of which includes the number of observations, number of variables, latitude and longitude of station, number of predictands, and starting and ending dates of the dataset. The second record gives a description of the dataset in alphanumeric form. The third and subsequent records contain the data, one event for each record. The program reads the data from a disc file, one record at a time, all predictors for one event.

b. Card image input - by means of the card image input, the user controls the number and specification of the predictand categories and the variables offered for screening. Two levels of predictor formulation are available, the raw predictor set contained in the database and a set of up to 50 predictors derived from the raw predictors. The specifications of the derived predictors are controlled by the card image input, described in the user guide, and any combination of raw and derived predictors may be offered, subject to the limitations described below. Appendix B is a list of the available compositions for the derived predictors. When choosing derived predictors and assigning predictors to be screened, care must be exercised to avoid offering linearly dependent predictors. For example, if two variables <u>and</u> their sum (function 3) are all offered, the screening program may select any two of the three without difficulty. On the next screening sequence, however, the program will fail when the third predictor is tried due to the linear dependence. Also, when the discontinuous linear and binary deviations are used, the threshold must be chosen so that not all the derived values are zero. If all the values of any predictor are zero, the screening program will fail on the first cycle.

3.2 Limitations on input

Efforts have been made to keep the dimensions of the variables consistent throughout the program and all subroutines to facilitate changes if desired. The limitations are as follows:

1. All variables dimensioned (150) are used to handle input data and derived predictor values. The total set of input and derived predictors must not exceed 150.

2. All variables dimensioned (50) are associated with input of derived predictors. The number of derived variables cannot exceed 50.

3. All variables dimensioned (100) are used to handle the full dataset that is input to screening. No more than 100 variables may be offered for screening.

4. All variables with dimension (10) are used to carry information concerning the predictors which have been selected by the screening program. Screening stops when 10 predictors have been selected even if the cutoff criterion has not yet been reached.

5. All variables with dimension (6) are used to carry information for the different categories. A maximum of 6 categories is permitted.

6. At present, a maximum of 200 events may be handled. This can be enlarged by increasing the dimension of ARRAY throughout the program and subroutines. ARRAY carries the data, reorganised into categories, which is offered to the MDA. The dimension of the predictand array Y should also be changed.

3.3 Major Steps of the Program

A. Main Program

1. Read data control variable from disc and read alphanumeric record. Read in category definitions.

2. Read in data, store predictand values, and count category sizes. Check that sample size in each category is reasonable. If any category has less than two events, the program stops.

3. Rewind dataset, read over first two records. Read card input data to end of predictor modification cards. Print out changes as requested. Read sequence numbers of variables to be included in MDA, print them out (MINDV).

4. Main loop for making changes as requested and selecting predictors to be offered. Events are read from disc, one at a time, changes are made, and requested predictors are selected. Each event is placed in array ARRAY such that all category 1 events are first, category 2 events next, and so on.

5. Call the screening program SDISCR. ARRAY is input and the program returns the index numbers of the predictors chosen, the number of predictors chosen, the within and between group matrices for predictors chosen, and the matrix product $(W^{-1}B)$. SDISCR uses an IMSL matrix inversion routine. Step-by-step statistics on the screening are printed by the screening program.

6. Print out diagnostic information:

   - Sample size, group sizes.
   - Actual (input) index numbers of predictors chosen. Note that the index numbers returned by the screening program are according to the sequence of predictors offered, from 1 to the number of predictors offered.
   - Climatological probabilities.
   - Group means for predictors chosen.
   - Grand means for predictors chosen.
   - Dependent sample predictor values for all events, for predictors chosen.

7. Calculate Eigenvalues and eigenvectors of $(W^{-1}B)$, using the IMSL routine EIGRF. Eigenvalues are returned in WEIG and eigenvectors are in ZZ.

8. Calculate and print out statistical test information on the eigenvalues.

9. Check for collinearity of group means. If three or more of the group means lie on or near a straight line, the number of independent discriminant functions is reduced. Collinearity is indicated by a very small eigenvalue. This is tested by comparing adjacent eigenvalues in the set. If the ratio drops below .001, all subsequent eigenvalues are rejected.

   At this point, the discriminant functions have been obtained for the predictors screened into the equation. The coefficients are the eigenvectors. To save space, the program from this point on overwrites data in the arrays used up to this point.

10. Convert the dependent sample to discriminant space. Subroutine CONVERT is called. ARRAY and the screening index numbers are input, and the program replaces the first IDM rows of ARRAY with the corresponding discriminant function values where IDM is the number of discriminant functions used.

11. Calculate the pooled within-group dispersion matrix for the dependent sample in discriminant space and invert it. Program WBMATX is used to calculate the

dispersion matrix and the discriminant functions' means, and the IMSL routine LINVIF is called to invert the matrix.

At this point, all the output parameters of the MDA that are needed for application to independent data have been calculated. They are:

a. The coefficients of the discriminant functions. (An array of size number of predictors by number of functions.)

b. The dependent sample means of the discriminant functions. (An array of size number of groups by number of funtions.)

c. The estimated climatological probabilities from the dependent sample. (A vector of length number of groups.)

d. The inverse of the pooled within groups dispersion matrix for the dependent sample in discriminant space. (An array of size number of functions by number of functions.)

From this point on, the program carries out calculations on the dependent data similar to those required in applications to independent data.

12. Calculate forecast probabilities for the dependent sample. Subroutine PROB is called within the loop with one set of predictor values, and all the dependent sample information outlined in (11) above. The output is a vector of group probabilities.

13. Calculate dependent sample diagnostics:

- Brier score
- Total variance (climate Brier score)
- Reduction of variance

Print out forecast probabilities and diagnostics.

B. Subroutine SDISCR

To determine the best predictors for discriminating among groups, the Mahalanobis D^2 statistic must be computed. This involves calculation of the within- and between-groups' dispersion matrices and inversion of the within-group's matrix. Much of the discriminant analysis must be done simply to select the best predictors, and the screening cannot therefore be separated completely from the discriminant analysis if maximum efficiency is to be achieved.

The screening procedure is a forward-selection method in the sense that a predictor that is entered remains for all subsequent steps. The main steps of the program are:

1. Calculate the within and between groups' dispersion matrices for <u>all</u> candidate predictors, calculate the group means and grand means for all candidate predictors.

2. Begin screening. Enter all predictors one at a time, calculate the Mahalanobis statistic for each, choose as first predictor the one with the highest value.

3. Search all predictors for the one which produces the highest Mahalanobis value in combination with the predictors already in the equation. To do this, the appropriate within and between groups' matrix elements are extracted from the matrices calculated in (1) above, and formed into smaller W and B matrices. This step is repeated until no predictor increases the Mahalanobis value by more than 10%. As each predictor test involves inversion of a matrix, computation time can be saved by restricting the number of predictors offered and/or restricting the number of predictors to be screened in. It is this part of the program that consumes most of the time.

At the conclusion of the screening, the output matrices W and B contain the within- and between- groups' dispersion for the selected predictors.

The screening routine provides for forcing of predictors into the equation. If a predictor is to be forced in, the corresponding element of input vector IDX should be set to: 1 to force the variable in
2 to delete the variable from consideration.
IDX is defined in a data statement in the main program.

C. Subroutine WBMATX

This program calculates within-group and between-group dispersion matrices, group means and grand means. Input data is in array X, output matrices are in XW and XB, group means are in XBAR and grand means are in XGM. This program is used twice: once to calculate dispersion matrices and means for all the predictors prior to screening, and once to calculate the within group's matrix and means for the discriminant functions for the dependent sample.

D. Subroutine CONVERT

This program transforms an input array of up to 100 predictors by 200 events to discriminant space, using coefficients from array COEF. The transformed data are put back into the input array in the first IDM rows where IDM is the number of discriminant functions.

E. Subroutine PROB

This program does all necessary computations to obtain forecast probabilities from discriminant functions. As input, it expects all of the MDA output parameters listed after (11) above, and the predictor values. A switch is provided to indicate whether the predictor values have been transformed to discriminant space. The program first computes discriminant functions from the predictor values if necessary, then determines deviations of the discriminant function values from the mean of each group (DEV(J)). Then the vector DEV pre- and post- multiplies the matrix W^{-1} for the discriminant functions to obtain the exponent of the normal distribution (estimated). Finally, Bayes' rule is applied to obtain the probabilities of membership in each group.

F. Subroutine BRIER

This program calculates the Brier score (Brier, 1950) given a vector of probabilities for NG groups and a verifying observation value (OBS). If the verifying category number is known, it is input in IGR. If not, IGR is set to 0 on input and the program calculates the verifying category. The Brier score calculated is added to SUM which must be predefined on input.

G. Subroutine PRINTPR

This program prints out the plain language description of the predictors offered to the screening program. The input data (ITEXT) has been read from the second record of the data file.

4. SAMPLE OUTPUT OF PROGRAM

The example shown below was set up for forecasting probability of precipitation amount for Hannover, Germany in 3 categories, less than 0.5mm in 24 hours, 0.5mm to 5.0mm in 24 hours and more than 5.0mm in 24 hours. The dataset accessed by the program contained 74 events during the winter of 1980-81. Model output parameters were chosen for inclusion in the database specifically for the precipitation forecasting problem, then the raw parameters were composited in various ways to form a final set of potential predictors that should be physically related to the occurrence of precipitation.

The various components of the program output are illustrated and discussed below:

A. Database description

The user-supplied heading is printed, followed by number of events, number of raw predictors, latitude and longitude of the station and database starting and ending dates. The last six items are read from the first record of the database file.

```
MDA-TESTRUN ON HANNOVER DATA WITH NEW FUNCTIONS AND NEW DATA
NO. OF EVENTS IS  74
NO OF INPUT PREDICTORS IS 100
STN COORDS ARE    52.47    9.70
START AND END DATES REQUESTED ARE 801209 810228
```

B. Derived predictor compositions

Each composition is listed, first in the form supplied on input, then in plain language form, where the new predictor's sequence number is also identified. The new predictors are added to the original input set, and sequencing starts with the next integer. The example shown here contains the maximum 50 compositions and demonstrates that it is possible to submit a second and higher level of composition. For example, to obtain positive vorticity squared, the two components of vorticity, $\frac{\partial v}{\partial x}$ and $\frac{\partial u}{\partial y}$ were combined using the subtraction function, then the resultant predictors are passed through the discontinuous linear function to set all negative vorticity values to 0, then the result is squared to produce a function which is constant for all negative values of vorticity and steeply rising for positive values. Through the use of such functions, it is possible to include non-linear effects in the linear statistical screening procedure.

```
 4  60  70  0.0000  0.0000
INPUT VARIABLE  101 IS THE DIFFERENCE,INPUT VARIABLE   60 MINUS INPUT VARIABLE   70

 4  61  71  0.0000  0.0000
INPUT VARIABLE  102 IS THE DIFFERENCE,INPUT VARIABLE   61 MINUS INPUT VARIABLE   71

 4  62  72  0.0000  0.0000
INPUT VARIABLE  103 IS THE DIFFERENCE,INPUT VARIABLE   62 MINUS INPUT VARIABLE   72

 4  65  75  0.0000  0.0000
INPUT VARIABLE  104 IS THE DIFFERENCE,INPUT VARIABLE   65 MINUS INPUT VARIABLE   75

 4  66  76  0.0000  0.0000
INPUT VARIABLE  105 IS THE DIFFERENCE,INPUT VARIABLE   66 MINUS INPUT VARIABLE   76

 4  67  77  0.0000  0.0000
INPUT VARIABLE  106 IS THE DIFFERENCE,INPUT VARIABLE   67 MINUS INPUT VARIABLE   77

 2 101   0  0.0000 99.9999
INPUT VARIABLE  107 IS THE EXCESS OF INPUT VARIABLE  101 OVER  0.0000, OTHERWISE ZERO.

 2 102   0  0.0000 99.9999
INPUT VARIABLE  108 IS THE EXCESS OF INPUT VARIABLE  102 OVER  0.0000, OTHERWISE ZERO.

 2 103   0  0.0000 99.9999
INPUT VARIABLE  109 IS THE EXCESS OF INPUT VARIABLE  103 OVER  0.0000, OTHERWISE ZERO.

 2 104   0  0.0000 99.9999
INPUT VARIABLE  110 IS THE EXCESS OF INPUT VARIABLE  104 OVER  0.0000, OTHERWISE ZERO.

 2 105   0  0.0000 99.9999
INPUT VARIABLE  111 IS THE EXCESS OF INPUT VARIABLE  105 OVER  0.0000, OTHERWISE ZERO.

 2 106   0  0.0000 99.9999
INPUT VARIABLE  112 IS THE EXCESS OF INPUT VARIABLE  106 OVER  0.0000, OTHERWISE ZERO.

13 107   0  0.0000  2.0000
INPUT VARIABLE  113 IS VARIABLE  107 PLUS CONSTANT    0.00 TAKEN TO THE POWER  2.00

13 108   0  0.0000  2.0000
INPUT VARIABLE  114 IS VARIABLE  108 PLUS CONSTANT    0.00 TAKEN TO THE POWER  2.00

13 109   0  0.0000  2.0000
INPUT VARIABLE  115 IS VARIABLE  109 PLUS CONSTANT    0.00 TAKEN TO THE POWER  2.00

13 110   0  0.0000  2.0000
INPUT VARIABLE  116 IS VARIABLE  110 PLUS CONSTANT    0.00 TAKEN TO THE POWER  2.00

13 111   0  0.0000  2.0000
INPUT VARIABLE  117 IS VARIABLE  111 PLUS CONSTANT    0.00 TAKEN TO THE POWER  2.00

13 112   0  0.0000  2.0000
INPUT VARIABLE  118 IS VARIABLE  112 PLUS CONSTANT    0.00 TAKEN TO THE POWER  2.00

 4  20  25  0.0000  0.0000
INPUT VARIABLE  119 IS THE DIFFERENCE,INPUT VARIABLE   20 MINUS INPUT VARIABLE   25

 4  21  26  0.0000  0.0000
INPUT VARIABLE  120 IS THE DIFFERENCE,INPUT VARIABLE   21 MINUS INPUT VARIABLE   26

 4  22  27  0.0000  0.0000
INPUT VARIABLE  121 IS THE DIFFERENCE,INPUT VARIABLE   22 MINUS INPUT VARIABLE   27

 2 119   0 30.0000 99.9999
INPUT VARIABLE  122 IS THE EXCESS OF INPUT VARIABLE  119 OVER 30.0000, OTHERWISE ZERO.

 2 120   0 30.0000 99.9999
INPUT VARIABLE  123 IS THE EXCESS OF INPUT VARIABLE  120 OVER 30.0000, OTHERWISE ZERO.

 2 121   0 30.0000 99.9999
INPUT VARIABLE  124 IS THE EXCESS OF INPUT VARIABLE  121 OVER 30.0000, OTHERWISE ZERO.

 2  15   0  0.0000 99.9999
INPUT VARIABLE  125 IS THE EXCESS OF INPUT VARIABLE   15 OVER  0.0000, OTHERWISE ZERO.

 2  16   0  0.0000 99.9999
INPUT VARIABLE  126 IS THE EXCESS OF INPUT VARIABLE   16 OVER  0.0000, OTHERWISE ZERO.

 2  17   0  0.0000 99.9999
INPUT VARIABLE  127 IS THE EXCESS OF INPUT VARIABLE   17 OVER  0.0000, OTHERWISE ZERO.

 5  30  30  0.0000  0.0000
INPUT VARIABLE  128 IS THE PRODUCT, INPUT VARIABLE   30 TIMES INPUT VARIABLE   30

 5  31  31  0.0000  0.0000
INPUT VARIABLE  129 IS THE PRODUCT, INPUT VARIABLE   31 TIMES INPUT VARIABLE   31

 5  32  32  0.0000  0.0000
INPUT VARIABLE  130 IS THE PRODUCT, INPUT VARIABLE   32 TIMES INPUT VARIABLE   32

 5  35  35  0.0000  0.0000
INPUT VARIABLE  131 IS THE PRODUCT, INPUT VARIABLE   35 TIMES INPUT VARIABLE   35

 5  36  36  0.0000  0.0000
INPUT VARIABLE  132 IS THE PRODUCT, INPUT VARIABLE   36 TIMES INPUT VARIABLE   36

 5  37  37  0.0000  0.0000
INPUT VARIABLE  133 IS THE PRODUCT, INPUT VARIABLE   37 TIMES INPUT VARIABLE   37

 3 128 131  0.0000  0.0000
INPUT VARIABLE  134 IS THE SUM OF INPUT VARIABLE  128 AND INPUT VARIABLE  131

 3 129 132  0.0000  0.0000
INPUT VARIABLE  135 IS THE SUM OF INPUT VARIABLE  129 AND INPUT VARIABLE  132

 3 130 133  0.0000  0.0000
INPUT VARIABLE  136 IS THE SUM OF INPUT VARIABLE  130 AND INPUT VARIABLE  133

11  30  35  0.0000  0.0000
INPUT VARIABLE  137 IS THE SQUARE ROOT OF THE SUM OF SQUARES OF VARIABLES   30 AND   35

11  31  36  0.0000  0.0000
INPUT VARIABLE  138 IS THE SQUARE ROOT OF THE SUM OF SQUARES OF VARIABLES   31 AND   36

11  32  37  0.0000  0.0000
INPUT VARIABLE  139 IS THE SQUARE ROOT OF THE SUM OF SQUARES OF VARIABLES   32 AND   37

 5 137  55  0.0000  0.0000
INPUT VARIABLE  140 IS THE PRODUCT, INPUT VARIABLE  137 TIMES INPUT VARIABLE   55

 5 138  56  0.0000  0.0000
INPUT VARIABLE  141 IS THE PRODUCT, INPUT VARIABLE  138 TIMES INPUT VARIABLE   56

 5 139  57  0.0000  0.0000
INPUT VARIABLE  142 IS THE PRODUCT, INPUT VARIABLE  139 TIMES INPUT VARIABLE   57

 5 137 119  0.0000  0.0000
INPUT VARIABLE  143 IS THE PRODUCT, INPUT VARIABLE  137 TIMES INPUT VARIABLE  119

 5 138 120  0.0000  0.0000
INPUT VARIABLE  144 IS THE PRODUCT, INPUT VARIABLE  138 TIMES INPUT VARIABLE  120

 5 139 121  0.0000  0.0000
INPUT VARIABLE  145 IS THE PRODUCT, INPUT VARIABLE  139 TIMES INPUT VARIABLE  121

 4   6  11  0.0000  0.0000
INPUT VARIABLE  146 IS THE DIFFERENCE,INPUT VARIABLE    6 MINUS INPUT VARIABLE   11

 4   7  12  0.0000  0.0000
INPUT VARIABLE  147 IS THE DIFFERENCE,INPUT VARIABLE    7 MINUS INPUT VARIABLE   12

 1   1   0-10.0000   .5000
INPUT VARIABLE  148 IS 1 WHEN INPUT VARIABLE    1 IS BETWEEN-10.0000 AND   .5000, INCLUSIVE, OTHERWISE ZERO.

 1   1   0   .5001  5.0000
INPUT VARIABLE  149 IS 1 WHEN INPUT VARIABLE    1 IS BETWEEN   .5001 AND  5.0000, INCLUSIVE, OTHERWISE ZERO.

 1   1   0  5.0000 99.9999
INPUT VARIABLE  150 IS 1 WHEN INPUT VARIABLE    1 IS BETWEEN  5.0000 AND 99.9999, INCLUSIVE, OTHERWISE ZERO.
```

C. List of predictors to be offered to the MDA

It is not necessary to offer all the available raw and derived predictors to the screening program. Predictors are selected for screening by the user by listing the sequence numbers of original and derived predictors wanted, 18 to a card in the input stream. This input is simply printed at this point for checking purposes.

```
  3   5  10  11  12  15  16  17  20  21  22  30  31  32  35  36  37  40
 41  42  45  46  47  50  51  52  55  56  57  60  61  62  65  66  67  80
 81  82  85  86  87  90  91  92  95  96  97 101 102 103 104 105 106 107
108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125
126 127 128 129 130 134 135 136 137 138 139 140 141 142 143 144 145 146
147
```

D. List of candidate predictors

Once the predictor selection has been made via the input stream (C. above), only those predictors that are to be offered are listed in plain language form. For the raw predictors, the plain language description has been taken from the second record of the dataset file. For the derived predictors, this information is not available and the reader is referred to the preceding composition descriptions.

CANDIDATE PREDICTORS ARE AS FOLLOWS

LIST OF PREDICTORS

NO.	PARAMETER	LEVEL	TIME STEP	CENTRL TIME	AREA AVERAGING	AREA INTERPOL.	DERI-VATION	TIME MEAN	TREND	
					AREA MANIPULATIONS			TIME MANIPULATIONS		
3	PRECIP. OB	0	96	96	-----	-----	-----	2 STEPS	TREND	PERSIST.
5	GEOPOT.	1000	72	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
10	GEOPOT.	500	72	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
11	GEOPOT.	500	84	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
12	GEOPOT.	500	96	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
15	GEOPOT.	1000	72	96	NO AVERAG.	NO INTERP	LAPLACIAN	1 STEPS	NO TREND	PREDICTOR
16	GEOPOT.	1000	84	96	NO AVERAG.	NO INTERP	LAPLACIAN	1 STEPS	NO TREND	PREDICTOR
17	GEOPOT.	1000	96	96	NO AVERAG.	NO INTERP	LAPLACIAN	1 STEPS	NO TREND	PREDICTOR
20	TEMPERAT.	1000	72	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
21	TEMPERAT.	1000	84	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
22	TEMPERAT.	1000	96	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
30	U-VELOCIT.	850	72	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
31	U-VELOCIT.	850	84	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
32	U-VELOCIT.	850	96	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
35	V-VELOCIT.	850	72	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
36	V-VELOCIT.	850	84	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
37	V-VELOCIT.	850	96	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
40	MIXING.R.	850	72	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
41	MIXING.R.	850	84	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
42	MIXING.R.	850	96	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
45	VERTI.WIND	700	72	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
46	VERTI.WIND	700	84	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
47	VERTI.WIND	700	96	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
50	VERTI.WIND	500	72	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
51	VERTI.WIND	500	84	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
52	VERTI.WIND	500	96	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
55	CLOUD COV	220	72	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
56	CLOUD COV	220	84	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
57	CLOUD COV	220	96	96	4POINTS	NO INTERP	NO DERIV.	3 STEPS	NO TREND	PREDICTOR
60	V-VELOCIT.	1000	72	96	4POINTS	NO INTERP	W-E GRADI.	3 STEPS	NO TREND	PREDICTOR
61	V-VELOCIT.	1000	84	96	4POINTS	NO INTERP	W-E GRADI.	3 STEPS	NO TREND	PREDICTOR
62	V-VELOCIT.	1000	96	96	4POINTS	NO INTERP	W-E GRADI.	3 STEPS	NO TREND	PREDICTOR
65	V-VELOCIT.	500	72	96	4POINTS	NO INTERP	W-E GRADI.	3 STEPS	NO TREND	PREDICTOR
66	V-VELOCIT.	500	84	96	4POINTS	NO INTERP	W-E GRADI.	3 STEPS	NO TREND	PREDICTOR
67	V-VELOCIT.	500	96	96	4POINTS	NO INTERP	W-E GRADI.	3 STEPS	NO TREND	PREDICTOR
80	TEMPERAT.	850	72	96	4POINTS	NO INTERP	W-E GRADI.	3 STEPS	NO TREND	PREDICTOR
81	TEMPERAT.	850	84	96	4POINTS	NO INTERP	W-E GRADI.	3 STEPS	NO TREND	PREDICTOR
82	TEMPERAT.	850	96	96	4POINTS	NO INTERP	W-E GRADI.	3 STEPS	NO TREND	PREDICTOR
85	TEMPERAT.	850	72	96	4POINTS	NO INTERP	N-S GRAD.	3 STEPS	NO TREND	PREDICTOR
86	TEMPERAT.	850	84	96	4POINTS	NO INTERP	N-S GRAD.	3 STEPS	NO TREND	PREDICTOR
87	TEMPERAT.	850	96	96	4POINTS	NO INTERP	N-S GRAD.	3 STEPS	NO TREND	PREDICTOR
90	LARGERAIN	220	84	100	4POINTS	NO INTERP	NO DERIV.	2 STEPS	TREND	PREDICTOR
91	LARGERAIN	220	96	100	4POINTS	NO INTERP	NO DERIV.	2 STEPS	TREND	PREDICTOR
92	LARGERAIN	220	100	100	4POINTS	NO INTERP	NO DERIV.	2 STEPS	TREND	PREDICTOR
95	CONVECRAIN	220	84	100	4POINTS	NO INTERP	NO DERIV.	2 STEPS	TREND	PREDICTOR
96	CONVECRAIN	220	96	100	4POINTS	NO INTERP	NO DERIV.	2 STEPS	TREND	PREDICTOR
97	CONVECRAIN	220	100	100	4POINTS	NO INTERP	NO DERIV.	2 STEPS	TREND	PREDICTOR
101	DERIVED PREDICTOR-SEE DESCRIPTION ABOVE									
102	DERIVED PREDICTOR-SEE DESCRIPTION ABOVE									
103	DERIVED PREDICTOR-SEE DESCRIPTION ABOVE									
104	DERIVED PREDICTOR-SEE DESCRIPTION ABOVE									
105	DERIVED PREDICTOR-SEE DESCRIPTION ABOVE									
106	DERIVED PREDICTOR-SEE DESCRIPTION ABOVE									
107	DERIVED PREDICTOR-SEE DESCRIPTION ABOVE									
108	DERIVED PREDICTOR-SEE DESCRIPTION ABOVE									
109	DERIVED PREDICTOR-SEE DESCRIPTION ABOVE									
110	DERIVED PREDICTOR-SEE DESCRIPTION ABOVE									
111	DERIVED PREDICTOR-SEE DESCRIPTION ABOVE									
112	DERIVED PREDICTOR-SEE DESCRIPTION ABOVE									
113	DERIVED PREDICTOR-SEE DESCRIPTION ABOVE									
114	DERIVED PREDICTOR-SEE DESCRIPTION ABOVE									
115	DERIVED PREDICTOR-SEE DESCRIPTION ABOVE									
116	DERIVED PREDICTOR-SEE DESCRIPTION ABOVE									
117	DERIVED PREDICTOR-SEE DESCRIPTION ABOVE									
118	DERIVED PREDICTOR-SEE DESCRIPTION ABOVE									
119	DERIVED PREDICTOR-SEE DESCRIPTION ABOVE									
120	DERIVED PREDICTOR-SEE DESCRIPTION ABOVE									

E. Screening Diagnostics

The next section of output describes the results of the screening. For the purposes of screening, all predictors offered are resequenced according to the order given in C. above. For example, predictor 40 is the 18th in the list and is therefore identified as number 18 by the screening program. The first two lines of output give the specifications for deletion of predictors and forcing of predictors, as controlled by the IDX parameter described above. At each step of screening, the best predictor is listed, along with the Mahalanobis test statistic value associated with the best predictor and the increment in the test statistic from the previous step. In the example shown, four predictors were selected, and none of the others produced an increase of at least 10% in the value of the test statistic. As the Mahalanobis statistic contains the sample size, the Mahalanobis values obtained, and therefore the Mahalanobis value increment required for discontinuing selection depends on the sample size. A final Mahalanobis value approaching 1 for every case in the sample (74 in the example) represents quite a good fit to the data and useful discriminant power. For larger samples, it is worthwhile to lower (relax) the cutoff criterion to 5% for two reasons: firstly, the larger sample provides greater confidence in the fit and can withstand the addition of more predictors that would result from a relaxed cutoff criterion; secondly, a larger sample size means generally larger Mahalanobis values and increased difficulty for new predictors to achieve a given percentage increase in the statistic. If the cutoff criterion is set too high, useful discriminant information will be omitted from the analysis (underfit) and, if it is set too low, unreliable predictors may be chosen, thereby reducing the confidence in the fit by fitting noise which is specific to the dependent sample (overfitting). The cutoff criterion is set by the variable PCT in the screening program, in a statement near the end of the program.

Following the screening information, the sample sizes are printed by category, the predicted variable number is given for checking purposes and the selected predictors are listed using the original sequence numbers and the plain language descriptions from the second record of the data file.

```
NUMBER OF VARIABLES DELETED     0
NO VARIABLES FORCED INTO ANALYSIS

FOR  1 PREDICTORS, THE BEST PREDICTORS ARE   5
MAX MAHALANOBIS D-SQUARE IS         37.44825
DIFFERENCE D-SQUARE( 1) - D-SQUARE( 0) IS        37.44825

FOR  2 PREDICTORS, THE BEST PREDICTORS ARE   5  85
MAX MAHALANOBIS D-SQUARE IS         53.57201
DIFFERENCE D-SQUARE( 2) - D-SQUARE( 1) IS        16.12376

FOR  3 PREDICTORS, THE BEST PREDICTORS ARE   5  85  70
MAX MAHALANOBIS D-SQUARE IS         62.19718
DIFFERENCE D-SQUARE( 3) - D-SQUARE( 2) IS         8.62517

FOR  4 PREDICTORS, THE BEST PREDICTORS ARE   5  85  70  47
MAX MAHALANOBIS D-SQUARE IS         69.28895
DIFFERENCE D-SQUARE( 4) - D-SQUARE( 3) IS         7.09178
ADDITIONAL PREDICTORS PRODUCE NO SIGNIFICANT IMPROVEMENT


TOTAL SAMPLE SIZE=  74 GROUP  1  37  GROUP  2  28  GROUP  3   9   GROUP

PREDICTAND IS VARIABLE NUMBER    1

PREDICTORS ARE
            12         141          123          97
:   12 : GEOPOT.    :    500  :     96  :     96  :    4POINTS  :  NO INTERP :  NO DERIV. :  3 STEPS   : NO TREND  PREDICTOR
---------------------------------------------------------------------------------------------------------------------------
    141 : DERIVED PREDICTOR-SEE DESCRIPTION ABOVE
---------------------------------------------------------------------------------------------------------------------------
    123 : DERIVED PREDICTOR-SEE DESCRIPTION ABOVE
---------------------------------------------------------------------------------------------------------------------------
:   97 : CONVECRAIN :    220  :    108  :    108  :    4POINTS  :  NO INTERP :  NO DERIV. :  2 STEPS   :   TREND   PREDICTOR
---------------------------------------------------------------------------------------------------------------------------
```

F. Dependent sample information

The climatological frequencies of the categories are given as estimated from the dependent sample. Then, the means of the predictors are listed by category. This information is very useful for a quick visual assessment of the analysis and a check that the selected predictors make sense from a meteorological point of view. For example, the first predictor chosen, 500mb height, shows the highest mean value for dry cases and lowest mean for category 3 (wettest cases), with a noticeable spread in the category means. Variable 141, the product of 850mb windspeed and cloud cover, increases from dry to wet cases, probably relating to the presence of frontal zones for heavier rain cases. Variable 123, a stability indicator in the form of 1000-500mb temperature difference in excess of 30^{o}C, also shows greater values for the precipitation cases than the dry cases but achieves highest mean values for the light rain cases. A possible explanation for this distribution of group means is that, in winter, heavy rain cases are associated with strong frontal (relatively stable) situations, while light rain cases are associated with weaker frontal situations or air mass convective activity. The fourth predictor, the model's convective precipitation, shows the expected relationship with the predictand.

The grand means are listed after the group means, followed by the predictor values for the complete dataset. These may be used to study cases of misclassification in the dependent sample, specifically to determine which predictors were responsible for the error.

CLIMATOLOGICAL FREQUENCIES (A PRIORI GROUP FREQUENCIES) ARE .5000 .3784 .1216

GROUP 1 MEANS

.5481E+04 .6814E+01 .1210E+01 .3739E+00

GROUP 2 MEANS

.5314E+04 .1169E+02 .4526E+01 .1452E+01

GROUP 3 MEANS

.5259E+04 .1466E+02 .2747E+01 .3132E+01

GRAND MEANS

.5391E+04 .9613E+01 .2652E+01 .1117E+01

PREDICTOR VALUES

	12	141	123	97
1	.5277E+04	.3695E+01	0.	.1431E-01
2	.5250E+04	.1054E+02	.1147E+01	.2284E+01
3	.5467E+04	.9674E+01	0.	.9155E-01
4	.5370E+04	.1305E+02	.2501E+01	.2080E+01
5	.5419E+04	.6473E+01	.9314E+01	.1379E+01
6	.5439E+04	.9144E+01	.3756E+01	.2747E+00
7	.5575E+04	.1097E+02	0.	.2670E-01
8	.5525E+04	.1861E+02	.8701E+00	.1216E+01
9	.5403E+04	.2240E+02	.2650E+01	.1602E+01
10	.5508E+04	.3388E+01	0.	.9537E-02
11	.5241E+04	.7515E+01	.9186E+01	.1107E+01
12	.5288E+04	.4592E+01	.1466E+01	.2480E+00
13	.5469E+04	.1123E+02	0.	.5674E-01
14	.5459E+04	.3069E+01	0.	.1907E-01
15	.5621E+04	.2282E+01	0.	0.
16	.5638E+04	.4046E+01	0.	.1431E-01
17	.5493E+04	.1001E+02	0.	.3328E+00
18	.5536E+04	.6696E+01	0.	.4768E-02
19	.5735E+04	.8150E+01	0.	0.
20	.5697E+04	.5817E+01	0.	.3471E+00
21	.5638E+04	.7656E+01	0.	.9289E+00
22	.5656E+04	.1441E+02	0.	.9632E-01
23	.5641E+04	.1273E+02	0.	.1278E+00
24	.5682E+04	.1156E+02	0.	0.
25	.5420E+04	.2749E+01	.3383E+01	.8392E-01
26	.5445E+04	.5098E+01	.6822E+00	.6771E-01
27	.5319E+04	.4176E+01	0.	.6771E+00
28	.5468E+04	.8082E+00	0.	.1049E-01
29	.5551E+04	.8734E+00	0.	-.2861E-02
30	.5546E+04	.3380E+00	0.	.9537E-03
31	.5487E+04	.2040E+00	0.	.4768E-02
32	.5374E+04	.4105E+01	.2743E+01	.1240E-01
33	.5417E+04	.2684E+01	.2154E+00	.2775E+00
34	.5334E+04	.1502E+01	.4520E+01	.2003E-01
35	.5326E+04	.1094E+01	.1816E+01	.2775E+00
36	.5557E+04	.5222E+01	0.	.6676E-02
37	.5516E+04	.5571E+01	.5108E+00	.1364E+00
38	.5282E+04	.1782E+02	.7437E+01	.9208E+00
39	.5485E+04	.7610E+01	.4463E+01	.1116E+00
40	.5327E+04	.6400E+01	.6338E+01	.1087E+00
41	.5213E+04	.1263E+02	.5811E+01	.5297E+01
42	.5293E+04	.1137E+02	.8383E+01	.3909E+01
43	.5306E+04	.2039E+02	.5891E+01	.2136E+01
44	.5092E+04	.1175E+02	.9758E+01	.2878E+01
45	.5162E+04	.7490E+01	.6382E+01	.1375E+01
46	.5341E+04	.6913E+01	.5042E+01	.4892E+00
47	.5426E+04	.1565E+02	0.	.9136E+00
48	.5106E+04	.1780E+02	.6703E+01	.4563E+01
49	.5144E+04	.1166E+02	.5626E+01	.2626E+01
50	.5201E+04	.1721E+02	.8322E+01	.3319E+01
51	.5257E+04	.8535E+01	.5403E+01	.1439E+01
52	.5285E+04	.9100E+01	.1566E+01	.1252E+01
53	.5428E+04	.1439E+02	0.	.3204E+00
54	.5536E+04	.1118E+02	0.	0.
55	.5337E+04	.1580E+02	.1207E+01	.1020E+00
56	.5105E+04	.1444E+02	.9784E+01	.1424E+01
57	.5184E+04	.1158E+02	.8871E+01	.2978E+01
58	.5305E+04	.2031E+02	.3625E+01	.9089E+00
59	.5356E+04	.1718E+02	.2055E+01	.1957E+01
60	.5422E+04	.1445E+02	.1988E+01	.2947E+00
61	.5406E+04	.2474E+01	.5270E+01	.5097E+00
62	.5451E+04	.8902E+01	.1581E+01	.2642E+00
63	.5385E+04	.5767E+01	.2258E+01	.1888E+00
64	.5467E+04	.3957E+01	.4648E+00	.3548E+00
65	.5503E+04	.4532E+01	.2501E+01	.4768E-02
66	.5364E+04	.1911E+02	.1518E+01	.3218E+01
67	.5386E+04	.9496E+01	.2917E+01	.3662E+00
68	.5340E+04	.2447E+02	.1289E+01	.3302E+01
69	.5255E+04	.1358E+02	.1037E+01	.7570E+01
70	.5125E+04	.1941E+02	.5874E+01	.3700E+01
71	.5256E+04	.1657E+02	.3509E+01	.3186E+01
72	.5054E+04	.1694E+02	.4692E+01	.5718E+01
73	.5238E+04	.1008E+02	.2042E+01	.3099E+00
74	.5314E+04	.2329E+01	.1844E+01	.8202E+00

G. Analysis results

Most of the information printed here is required for application to independent data-sets. The values themselves do not mean very much in meteorological terms, but some checks can be done on the analysis results, and some information can be gleaned regarding the relative importance of the discriminant functions.

The eigenvalues are listed in descending order of magnitude. The greater the difference in magnitude, the greater the difference in discriminant power of the associated discriminant functions. The following lines of output give the data required to perform Bartlett's test on the significance of the discriminant functions. The first is nearly always significant, the second and subsequent ones may be tested by comparing the given chi-square with values listed in chi-square tables for the given degrees of freedom. Further details of this and other tests are contained in Miller, 1962.

The MDA program automatically uses <u>all</u> the available functions in the dependent sample analyses that follow. If one or more should prove insignificant after testing, the program MDARUN can be used with the dependent data sample and a subset of the discriminant functions to recalculate the probabilities and other diagnostics.

The eigenvectors, the coefficients of the predictors in the discriminant functions, are listed in the same order as the order in which predictors were selected. The group means of the discriminant functions are calculated after transformation of the dependent data to discriminant space, then the within group dispersion matrix and its inverse are calculated for the data in discriminant space. The matrix should be diagonal to within roundoff error. An examination of the magnitude of the off-diagonal elements reveals arithmetic precision of the computations. If the off-diagonal elements rise much above 10^{-11} or so in magnitude (for a 60 bit computer word), this indicates computational problems, either a near-singular dispersion matrix caused by very highly or perfectly correlated input predictors, or nearly collinear group means.

```
EIGENVALUES        .8188E+00  .3171E+00

ROOT(1)=  .81881E+00     CHI-SQUARE  .41574E+02     DEGREES OF FREEDOM=  8

ROOT(2)=  .31707E+00     CHI-SQUARE  .19141E+02     DEGREES OF FREEDOM=  6


EIGENVECTORS


     1               -.2395E-01  .4128E+00 -.8884E-01  .8215E+00

     2                .1683E-02  .8775E-01  .7791E+00 -.1205E+01
MEANS OF DISCRIMINANT FUNCTIONS FOR DEPENDENT SAMPLE

  -.1283E+03  -.1217E+03  -.1176E+03

   .1031E+02   .1175E+02   .8502E+01


INVERSE OF POOLED WITHIN GROUP DISPERSION MATRIX IN DISCRIMINANT SPACE

   .49387E-01   .80796E-15

   .80796E-15   .28381E+00
```

H. Application output

The results of application of the discriminant analysis output to the dependent data-set are the last set of information printed. Probabilities are calculated for each of the events and are printed along with the actual predictand value and the group of its membership. As the MDA rearranges the input data in order of categories, this output is also organised according to categories, with all group one events followed by group 2 events, etc. This output can be quickly scanned to determine subjectively how good a fit was achieved with the dependent sample. Usually, the most interesting information is the accuracy of forecasts of the rarest category. In the given example, it can be seen that the MDA successfully fit 6 of the 9 category 3 events, and those that were not fit lay closest to the threshold with category 2. Like regression, MDA performs best on the dependent data, and the diagnostics presented here represent the best that is likely to be achieved.

Finally, the Brier Score (Brier, 1950) for a climatological forecast and the reduction of variance are printed. The Brier Score is actually the mean squared error of the probability forecast, compared with perfection as represented by an always correct categorical forecast. The Brier Score has a negative orientation, that is, smaller values (lower mean squared error) mean better forecasts. Best values are achieved by forecasting as close as possible to 100% probability for the category that occurs and as close as possible to 0% probability for the categories which do not occur. Because the errors are squared, a forecast of close to 100% probability for a category which <u>does not</u> occur is given a relatively high penalty. The climate score is the Brier Score that would be obtained if the climatological probabilities as estimated from the dependent sample were entered for every event. The reduction of variance represents the percentage improvement in the MDA forecasts over the climate forecast used as a standard of comparison and has the standard skill score form.

$$1 - \frac{\text{Score(F)}}{\text{Score(C)}}$$

where Score(F) is the Brier Score for the forecast and Score(C) is the Brier Score for the standard of comparison (climatology). This skill score is exactly analogous to the reduction of variance used as a measure of goodness of fit in regression.

A Brier Score under 0.20 represents a quite useful product, especially if there are more than 2 categories. Reduction of variances for probability forecasts over 25% represent a useful forecast, but will depend to some extent on the difficulty of the problem as represented by the original amount of variance given by the climate score. Climatology is a good predictor, especially when the categories are distributed such that one is very common and another is rare. In such cases, it is difficult to obtain an MDA which successfully forecasts the rare events without sacrificing skill in forecasting the common events and settling for a low reduction of variance. If the rare events are the most important from the meteorological point of view, a high <u>average</u> score can be sacrificed for extra skill in forecasting the rarer event.

PROBABILITY FORECASTS FOR DEPENDENT SAMPLE

EVENT	PDCTAND	ACT GROUP	FCST PROBABILITIES		
1	0.000	1	.681	.252	.067
2	.300	1	.193	.232	.576
3	.100	1	.778	.204	.018
4	.100	1	.382	.464	.154
5	.300	1	.299	.701	.000
6	0.000	1	.546	.451	.004
7	.400	1	.862	.132	.005
8	.200	1	.527	.416	.057
9	.100	1	.139	.759	.102
10	0.000	1	.932	.064	.004
11	0.000	1	.090	.909	.001
12	0.000	1	.608	.352	.040
13	0.000	1	.732	.247	.021
14	0.000	1	.910	.083	.007
15	0.000	1	.975	.025	.001
16	0.000	1	.970	.029	.001
17	.200	1	.807	.175	.018
18	0.000	1	.906	.089	.005
19	0.000	1	.970	.030	.000
20	0.000	1	.975	.024	.001
21	.200	1	.953	.043	.004
22	.100	1	.865	.131	.004
23	0.000	1	.885	.112	.003
24	.100	1	.925	.074	.001
25	.200	1	.769	.229	.002
26	.300	1	.847	.145	.009
27	.100	1	.725	.187	.088
28	0.000	1	.941	.055	.004
29	0.000	1	.967	.032	.001
30	.200	1	.969	.030	.001
31	.200	1	.953	.044	.003
32	.400	1	.691	.303	.006
33	.300	1	.884	.103	.013
34	0.000	1	.613	.384	.003
35	.200	1	.777	.208	.015
36	.100	1	.936	.062	.003
37	0.000	1	.901	.095	.004
38	1.400	2	.035	.962	.003
39	2.000	2	.631	.368	.001
40	1.000	2	.285	.713	.001
41	3.000	2	.038	.166	.797
42	1.400	2	.140	.801	.059
43	.700	2	.052	.907	.041
44	1.100	2	.020	.955	.025
45	2.000	2	.116	.860	.025
46	.600	2	.388	.607	.005
47	1.200	2	.483	.388	.129
48	2.200	2	.010	.334	.656
49	2.100	2	.070	.726	.205
50	2.800	2	.027	.912	.061
51	4.400	2	.224	.750	.026
52	1.100	2	.407	.426	.167
53	4.100	2	.543	.396	.061
54	4.100	2	.819	.172	.009
55	3.200	2	.259	.683	.059
56	1.100	2	.010	.986	.004
57	2.100	2	.051	.920	.029
58	1.400	2	.074	.882	.045
59	2.500	2	.222	.543	.236
60	2.500	2	.413	.567	.020
61	3.100	2	.669	.329	.001
62	2.400	2	.712	.278	.011
63	1.200	2	.682	.308	.010
64	4.000	2	.896	.096	.008
65	1.100	2	.849	.150	.001
66	8.100	3	.105	.219	.675
67	6.100	3	.499	.489	.012
68	11.200	3	.030	.168	.802
69	22.400	3	.001	.000	.999
70	10.200	3	.013	.477	.510
71	10.900	3	.076	.393	.531
72	11.200	3	.001	.019	.980
73	7.000	3	.270	.657	.074
74	7.700	3	.731	.234	.035

BRIER SCORE = .1857 CLIMATE SCORE= .2960 REDUCTION OF VARIANCE= .3726

REFERENCES

1. Brier, G.W., 1950: Verification of forecasts expressed in terms of probability, Monthly Weather Review, 79, 1-3.

2. Draper, N.R. and Smith, H., 1981: Applied regression analysis, second edition, Wiley & Sons, New York, p. 308.

3. Fisher, R.A., 1936: The use of multiple measurements in taxonomic problems, Annals of Eugenies, 7, part II, 179-188.

4. Miller, R.G., 1962: Statistical prediction by discriminant analysis, AMS Meteorological Monographs, 4, No. 25, 54 pp.

APPENDIX A Definitions of terms used in discriminant analysis

1. A priori probability

The unconditional probability of occurrence of one of the predictand groups. Usually refers to the natural frequency of occurrence of the predictand category, without any reference to predictor information. These are usually estimated from the frequencies of occurrence of the categories in the dependent sample.

2. Bayes' Rule

If there is a set of G mutually exclusive events that can occur, the probability of a particular event g, conditional on the a priori probabilities of the events q_g, (g=1,---G) and, given the knowledge of t predictors X_1---X_t, is:

$$P(g|X_1---X_t) = \frac{P(X_1---X_t|g).q_g}{\sum_{g=1}^{G} P(X_1---X_t|g).q_g}$$

3. Discriminant functions

A set of mutually uncorrelated (independent) linear combinations of a set of predictors which maximises the ratio of the between-group dispersion to the within-group dispersion for a given dataset.

4. Discriminant space

In MDA, the geometric counterpart to an observation is a point in space where the coordinates are the values of the discriminant functions associated with that point. Discriminant functions are the axes of the discriminant space.

5. Dispersion matrix

The term dispersion in this context refers to the spread or scatter of the observations from the mean value. In multivariate statistics, the dispersion matrix refers to the aggregate of variances and covariances of all variables. It is also referred to as the variance-covariance matrix, correlation matrix or matrix of sums of squares and crossproducts, depending on which statistic is used to measure the dispersion. In MDA, two dispersion matrices are used:

a. Within-groups dispersion matrix

The diagonal of this matrix is made up of variances of the predictors <u>with respect to the individual group means</u>, summed ("pooled") over all categories.

$$\sum_{g=1}^{G} \sum_{k=1}^{n_g} (X_{pgk} - \bar{X}_{pg})^2 \qquad p = 1, \text{---} P$$

where $\bar{X}_{pg}$ is the group g mean of predictor p.

The off-diagonal elements are cross-products of predictors, referenced to the group means, summed ("pooled") over all categories.

$$\sum_{g=1}^{G} \sum_{k=1}^{n_g} (X_{pgk} - \bar{X}_{pg})(X_{qgk} - \bar{X}_{qg}) \qquad p,q = 1, \text{---} P,\ p \neq q$$

This is the (p,q) element (and the (q,p) element since the matrix is symmetric). $\bar{X}_{pg}$ and $\bar{X}_{qg}$ are the group means of predictors p and q for category g.

b. Between-groups dispersion matrix

The diagonal of this matrix is made up of the variances of the predictor group means with respect to the overall ("grand") mean, summed over all groups.

$$\sum_{g=1}^{G} n_g (\bar{X}_{pg} - \bar{X}_p)^2 \qquad p = (1, \text{---} P)$$

where $\bar{X}_p$ is the overall sample mean of predictor p.

The off-diagonal elements are the crossproducts of the predictor group means referenced to the grand mean.

$$\sum_{g=1}^{G} n_g (\bar{X}_{pg} - \bar{X}_p)(\bar{X}_{qg} - \bar{X}_q) \qquad (p,q=1, \text{---} P)\ p \neq q$$

This is the (p,q) element (and q,p element). $\bar{X}_p$ and $\bar{X}_q$ are the grand means of predictors p and q.

These two matrices have the property that their sum is the total dispersion matrix that is used in regression analysis.

6. Grand mean

The mean of all the values of a predictor in the sample. This term is used in MDA to distinguish the sample mean from the group means which are means over only the values of the predictor in each specific group.

7. Multivariate normal distribution

The multivariate extension of the normal distribution, a frequency distribution of the vector variable $\overline{X}(\overline{X}=X_1 --- X_p)$

$$f(\overline{X}) = \frac{|\Sigma^{-1}|^{\frac{1}{2}}}{(2\pi)^{P/2}} \exp\left[-\tfrac{1}{2}(\overline{X}-\overline{\mu})'\ \Sigma^{-1}(\overline{X}-\overline{\mu})\right]$$

$$-\infty < X_1 < \infty$$
$$\vdots$$
$$-\infty < X_p < \infty$$

where $\overline{\mu}$ is the population mean vector

and Σ is the population variance-covariance matrix.

The multivariate normal distribution is a normal distribution in each of its dimensions.

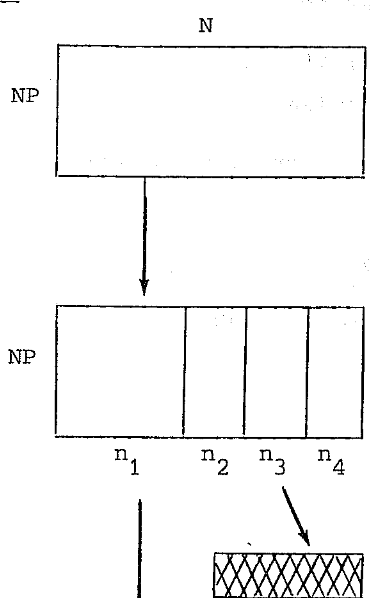
APPENDIX B Functions available to the MDA program

1.	Binary	$Y = 1$ if $X_1 \in (a,b)$ $Y = 0$ otherwise
2.	Discontinuous linear	$Y = (X_1 - a)$ if > 0 $Y = 0$ otherwise
3.	Sum of 2 predictors or 1 predictor and constant	$Y = (X_1 + X_2)$ or $Y = (X_1 + a)$
4.	Difference of 2 predictors	$Y = (X_1 - X_2)$
5.	Product of 2 predictors or 1 predictor and constant	$Y = X_1 X_2$ or $Y = aX_1$
6.	Ratio of 2 predictors	$Y = X_1 / X_2$
7.	Sin (day of year)	
8.	Cos (day of year)	
9.	Sin 2 (day of year)	
10.	Cos 2 (day of year)	
11.	SQRT of sum of squares of 2 variables	$Y = (X_1^2 + X_2^2)^{\frac{1}{2}}$
12.	Discontinuous linear (negative slope version)	$Y = a - X_1$ if > 0 $Y = 0$ otherwise
13.	Power	$Y = (X_1 \pm a)^b$
14.	Exponential	$Y = \exp(X_1 \pm a)$
15.	Exponential (negative argument)	$Y = \exp(-X_1 \pm a)$
16.	Natural logarithm	$Y = \ell n(X_1 \pm a)$

Where Y represents the derived predictor, X_1 and X_2 are the input raw predictors involved and a and b are constants specified on input.

APPENDIX C Data flow in MDA program

1. Input



a. Array of NP potential predictors by N events

b. G - Number of groups

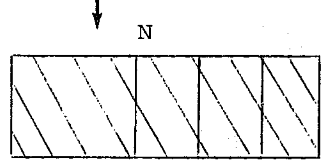
c. GB - Thresholds) to partition data

d. Predictand) into categories

Partitioned and reorganised into G sub-arrays of size NP x Ng, laid side-by-side in original array. Sample sizes are n_1, n_2---n_G and they must add up to N.

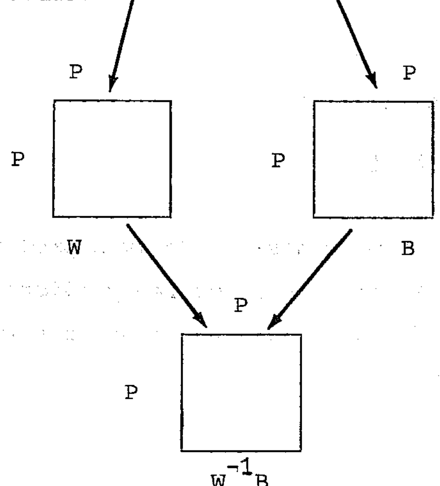
Group frequencies (1XG) vector

2. Screening



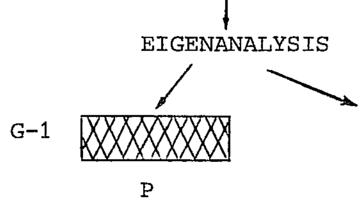
Selects P predictors, usually < 10 from the original NP. Now have PXN array, still partitioned according to predictand categories.

3. Analysis



Two dispersion matrices P X P are calculated from original partitioned dataset.

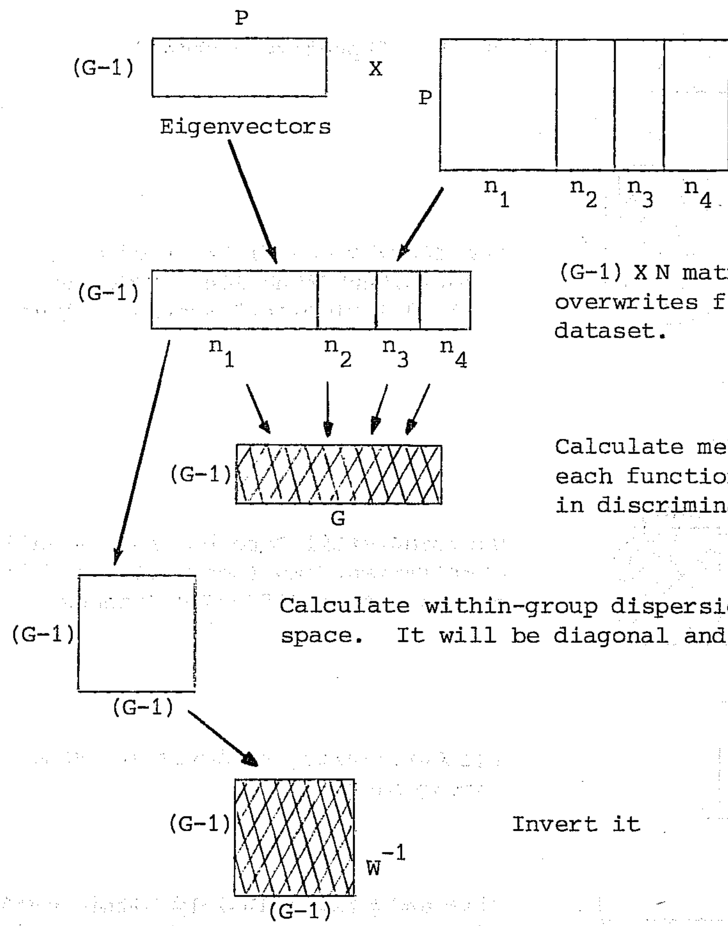
Invert W and pre-multiply B. Matrix is still P X P.



(G-1) eigenvalues
(G-1) eigenvectors of length P

[If P < G-1, P eigenvalues and P eigenvectors result. Usually, G-1 is < P]

4. <u>Data conversion</u>



Convert original P X N dataset to discriminant space

(G-1) X N matrix of transformed data overwrites first G-1 rows of original dataset.

Calculate means, one for each group, for each function → (G-1) X G array of means in discriminant space.

Calculate within-group dispersion matrix in discriminant space. It will be diagonal and (G-1) X (G-1)

Invert it

The program prints out arrays that are hatched. The cross-hatched arrays are printed <u>and</u> are required in application to dependent or independent datasets for calculation of probabilities.

5. Application