DECUS
PROGRAM LIBRARY

DECUS NO.	8-328
TITLE	NNAN (NEAREST NEIGHBOR ANALYSIS) - OREG (ORTHOGONALIZED REGRESSION ANALYSIS) - OREH (ADDITIONAL ORTHOGONAL REGRESSION COEFFICIENTS
AUTHOR	J. N. R. Jeffers
COMPANY	Submitted by: A. J. P. Gore The Nature Conservancy Grange-over-Sands Lancashire, England
DATE	September 26, 1969
SOURCE LANGUAGE	FORTRAN D

NNAN (NEAREST NEIGHBOR ANALYSIS)
OREG (ORTHOGONALIZED REGRESSION ANALYSIS)
OREH (ADDITIONAL ORTHOGONAL REGRESSION COEFFICIENTS)

DECUS Program Library Write-up

DECUS No. 8-328

NNAN (NEAREST NEIGHBOR ANALYSIS)

SUMMARY

This program computes, for each of a set of points, the nearest neighbor in n-dimensional space, and the distance to this neighbor. The program is intended for use with sets of data for which the number of dimensions has been reduced to a relatively small number of orthogonal axes, but the algorithm is applicable to other types of data.

TAPES REQUIRED

Form of program tape - The program is written in the PDP-8 FORTRAN-D language, and is in the source language.

Form of data tape - The data to be analyzed should be punched onto paper tape in the ASCII code. They should consist of the successive co-ordinates of the individuals, in a standard order. The points themselves will be given serial numbers by the program, but these identification numbers should not occur on the data tape. The output tapes from the CVAL and GRMN programs are suitable for input without modification.

OPERATING INSTRUCTIONS

```
.FORT                          Source program tape in high-speed reader
*OUT-S:NNAN
*
*IN-R:
* ↑
↑                              Data tape in high-speed reader
*READY
↑
```

The program will request the entry of the number of variables and data sets, and these should be entered on the teletype and terminated by "Return". If an output tape is required, switch on the low-speed punch before typing "Return".

If the program has already been compiled onto the disk, it may be called back into core as follows:

```
.FOSL
*IN-S:NNAN
*
*OPT-
* ↑                          Data tape in high-speed reader
*READY
↑
```

OUTPUT

The program prints, for each point in turn, the identification number of the point being considered, the identification of the nearest neighbor, and the distance between these two points. The output from this program is directly suitable for the CLAN program which carries out a cluster analysis.

STORAGE AND LIMITATIONS

Normal for FORTRAN-D For the 4K version, the number of variables times the number of sets must not exceed 260. This restriction can be relaxed by using the 8K version.

METHOD

The program computes the distance between the reference point and every other point in the set, and stores the shortest distance. The distance is computed from the formula:

$$\text{Distance} = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \ldots\ldots + (X_{ni} - X_{nj})^2}$$

```
*R

*L

C;      PROGRAM SEGMENT NNAN
        DIMENSION V(26Ø)
        TYPE 1ØØ
1ØØ     FORMAT (/, "ENTER NOS OF VARIABLES AND SETS ", /, /)
        ACCEPT 1Ø1, N1, N2
1Ø1     FORMAT (I, I)
        N3=N1*N2
        DO 2Ø I=1, N3
        READ 2, 1Ø3, V(I)
2Ø;     CONTINUE
1Ø3     FORMAT (E)
        N1Ø=1
        N4=Ø
4;      N5=Ø
        N8=1
        VØ=1ØØØØØØØØØ.Ø
3;      IF (N5-N4) 5, 1, 5
5;      N6=N4
        N7=N5
        NØ=Ø
        V1=Ø
2       V2=V(1+N6)-V(1+N7)
        V1=V1+(V2*V2)
        NØ=NØ+1
        N6=N6+1
        N7=N7+1
        IF (N1-NØ) 2, 6, 2
6;      V1=SQTF(V1)
        IF (V1-VØ) 7, 1, 1
7;      VØ=V1
        N9=N8
1       N5=N5+N1
        N8=N8+1
        IF (N2-N8) 8, 3, 3
8;      TYPE 1Ø2, N1Ø, N9, VØ
1Ø2     FORMAT (/, I, I, E)
        N1Ø=N1Ø+1
        N4=N4+N1
        IF (N2-N1Ø) 9, 4, 4
9;      STOP
        END

*
```

OREG (ORTHOGONALIZED REGRESSION CALCULATION)

SUMMARY

The program calculates the orthogonalized regression of one or more dependent variables on the principal components of up to 20 regressor variables. The basic data required are the correlation coefficients between the dependent variables and the regressor variables, and the latent roots and vectors of the correlation matrix of the regressor variables.

TAPES REQUIRED

Form of program tape - The program is written in the PDP-8 FORTRAN-D language, and is in the source language.

Form of data tape - The data to be analyzed should be punched onto paper tape in the ASCII code. The data should begin with the latent roots for as many components as are required. The latent roots should then be followed by the latent vectors corresponding to each root. Finally, the coefficients of the correlations with the original variables should be given for each of the dependent variables, e.g.

Latent roots

3.2470 1.2753 0.3859 0.0700 0.0218

Latent vectors (punched by rows)

-0.5212	0.3121	0.4753	0.3245	-0.5471
-0.0711	0.7090	0.0381	-0.6943	0.0935
0.4730	0.2161	0.8246	-0.2224	0.0105
0.5219	0.5942	0.0103	0.5801	0.1945
-0.4757	-0.0112	0.3050	0.1629	0.8087

Consumption of oil

-0.457 0.032 0.899 0.601 -0.710

OPERATING INSTRUCTIONS

If the program has not previously been compiled onto the disk:

```
.FORT
*OUT-S:OREG
*
*IN-R:                    Insert program tape in high-speed reader
* ↑
*READY                    Insert data tape in high-speed reader
↑
```

The program will request the entry of the numbers of variables, latent roots, and dependent variables (correlations). These should be entered on the teletype, and terminated by "return".

If the program has already been compiled onto the disk, it may be called into the core store as follows:

```
.FOSL
*IN-S:OREG
*
*OPT-
*READY                    Insert data tape in high-speed reader
↑
```

OUTPUT

The program first prints the proportions of the variability of the dependent variable accounted for by each of the components. These are followed by the orthogonalized regression coefficients for each of the components.

STORAGE AND LIMITATIONS

Normal for FORTRAN-D The maximum number of regressor variables must not exceed 20, and the number of components times the number of regressor variables must not exceed 100.

METHOD

The calculations follow closely the method described by Kendall in Chapter 5 of "A Course In Multivariate Analysis". The latent vectors are first transformed by dividing them by the square root of their sum of squares. The resulting s_i are multiplied by the corresponding correlation coefficients and the sum of the products divided by the latent root. The reduction of the variance due to the fitting of the components is calculated, and the coefficients of the orthogonalized regressions expressed in terms of the standardized variates.

```
L

C       PROGRAM TO COMPUTE ORTHOGONALIZED REGRESSIONS, OREG
        DIMENSION V(1ØØ), CORR(2Ø), PROP(2Ø), REGR(2Ø)
        TYPE 1Ø4
1Ø4     FORMAT (/,"ENTER NO OF VARIABLES, ROOTS, CORRELATIONS",/)
        ACCEPT 1Ø1,N,M,K
1Ø1     FORMAT (I,I,I)
        DO 1Ø I=1,M
        READ 2,1Ø2,PROP(I)
1Ø2     FORMAT (E)
1Ø      CONTINUE
        NM=N*M
        DO 2Ø I=1,NM
        READ 2,1Ø2,V(I)
2Ø      CONTINUE
        DO 4Ø I=1,M
        PSI=Ø.Ø
        DO 5Ø J=1,N
        L=J+N*(I-1)
        PSI=PSI+V(L)*V(L)
5Ø      CONTINUE
        PSI=SQTF(PSI)
        DO 4Ø J=1,N
        L=J+N*(I-1)
        V(L)=V(L)/PSI
4Ø      CONTINUE
        DO 3Ø IN=1,K
        DO 6Ø I=1,N
        READ 2,1Ø2,CORR(I)
        REGR(I)=Ø.Ø
6Ø      CONTINUE
        DO 7Ø I=1,M
        DO 7Ø J=1,N
        L=J+N*(I-1)
        REGR(I)=REGR(I)+CORR(J)*V(L)
7Ø      CONTINUE
        DO 8Ø I=1,M
        REGR(I)=REGR(I)/PROP(I)
        CORR(I)=REGR(I)*REGR(I)*PROP(I)
        TYPE 1Ø3,CORR(I)
1Ø3     FORMAT (/,E)
8Ø      CONTINUE
        J=Ø
        DO 9Ø I-1,M
        TYPE  1Ø5
```

```
1Ø5     FORMAT (/)
        DO 1ØØ L=1, N
        LL=J+L
        PSI=REGR(I)*V(LL)
        TYPE 1Ø3, PSI
1ØØ     CONTINUE
        J=J+N
9Ø      CONTINUE
3Ø      CONTINUE
        STOP
        END
```

*

OREH (PROGRAM TO ADD ORTHOGONALIZED REGRESSION COEFFICIENTS)

SUMMARY

This program is an auxiliary program to OREG, the orthogonalized regression program. It adds the corresponding orthogonalized regression coefficients for nominated components to give a single vector of standardized regression coefficients. Being standardized, these regression coefficients may be used to judge the relative importance of the original variables in estimating values of the dependent variable.

TAPES REQUIRED

Form of program tape - The program is written in the PDP-8 FORTRAN-D language, and is in the source language.

Form of data tape - The data tape for this program is the output tape from the OREG program. If an output tape was not obtained from the original run of this program, a special version of OREG enables these results to be punched on the high-speed punch.

METHOD OF OPERATION

Normal for the disk operating system. The data tape should be loaded into the high-speed reader before continuing after the teletype has printed READY. The program will request the entry of the number of variables and components and will pause after reading each set of orthogonalized regression coefficients. If the coefficients of that set are required to be added, type 1 (space). Otherwise, type 0 (space).

OUTPUT

The output from this program consists of the algebraic sum of the corresponding orthogonalized regression coefficients which have been nominated for inclusion by the operator.

STORAGE AND LIMITATIONS

Normal for FORTRAN-D. As for OREG, the number of initial variables is limited to 20.

```
L

C       PROGRAM TO ADD REGRESSION COEFFICIENTS
        DIMENSION X(2Ø), SX(2Ø)
        TYPE 1ØØ
1ØØ     FORMAT (/, "ENTER NO OF VARIABLES",/)
        ACCEPT 1Ø1, N
1Ø1     FORMAT (I)
        DO 1Ø I=1, N
        READ 2, 1Ø2, X(I)
        SX(I)=Ø.Ø
1Ø      CONTINUE
        DO 2Ø I=1, N
        DO 3Ø J=1, N
        READ 2, 1Ø2, X(J)
3Ø      CONTINUE
        ACCEPT 1Ø1, K
        IF (K) 1, 2Ø, 1
1       DO 4Ø J=1, N
        SX(J)=SX(J)+X(J)
4Ø      CONTINUE
2Ø      CONTINUE
        DO 5Ø I=1, N
        TYPE 1Ø2, SX(I)
5Ø      CONTINUE
1Ø2     FORMAT (/, E)
        STOP
        END

*
```