

Construction of Input for Environmental Simulation Models using Multivariate Analysis.

Joaquin Sanabria*, and Don W. Goss*

INTRODUCTION

With the present time storage capacity and speed of computers, simulation models have become one of the most powerful research tools in environmental studies. In many exploratory studies that require assessment of the effect of many factors and the interactions between factors on natural resources, simulation models are the only tool to cover extensive geographical areas. In detailed environmental studies with the main purpose of identifying participating factors and their roles, simulation models are the only methodological resource available to analyze innumerable possible results from also innumerable combination of factors, in relatively short time at low costs.

Three of the environmental simulation models better known and widely applied worldwide are EPIC (Environmental Policy Integrated Climate)(Williams, et al, 1990), APEX (Agricultural Policy/Environmental eXtender)(Williams, et al, 1996), and SWAT (Soil and Water Assessment Tool)(Arnold, et al, 1993). The three models are time continuous models used to determine the effect of soil, climate, watershed characteristics, and management strategies on soil and water resources. They are used to evaluate soil erosion, flow and sediment load in streams, losses of pesticides and nutrients from farmland fields and animal production sites, as well as transport and accumulation of agro-chemicals in streams lakes and other water bodies. EPIC and APEX are also used for estimation of agricultural production under a high diversity of conditions resulting from the combination of soil, climate and management factors. The three models have in common the same physical, and bio-chemical principles and differ mainly in the operation scale.. EPIC operates at field level, where the assumed homogeneous drainage area goes from one to few hectares. APEX operates at the level of landscape integrating the heterogeneous conditions of several sites. SWAT works at the level of watershed considering the variability of a drainage area from few to hundreds of hectares.

After a simulation model is validated and calibrated for a particular set of research conditions, the quality of its performance making realistic estimations is proportional to the quality of the input data. The input data for the three mentioned simulation models is composed by three major components: Climate data, soil data, and management data.

*Texas Agricultural Experiment Station
Blackland Research Center
Texas A&M University System

The objectives of this paper are two:

- 1) To use the multivariate methods cluster and factor analysis to obtain input data for environmental simulation models.
- 2) To demonstrate the use of Multivariate SAS procedures such as PROC FACTOR, PROC CLUSTER, and PROC FASTCLUS, and other SAS procedures like PROC STANDARD, and PROC SCORE in the construction of input data sets for environmental simulation models.

METHODOLOGY

This paper deals mainly with statistical methodologies for obtaining the climate and soil components of input for the simulation model EPIC, since this is the simulation model employed in the research project that originated this paper. Statistical methodologies for construction of management input are not considered here given that the number of possibilities of this type of input are few, compared with the overwhelming number of possible climate and soil entries for regional or national studies in the US.

CLIMATE INPUT

The model admits either entering real daily data or monthly parameters for weather generation. The daily data required are maximum temperature, minimum temperature, precipitation, solar radiation, relative humidity, and wind velocity, for the period of time that is going to be simulated. Finding the data described above for the specific periods of time that is going to be analyzed and for the location of interest is difficult, for that reason EPIC, APEX, and SWAT have weather generators able to build weather data sets for any location and period of time. Weather generator is also required in prediction of environmental factors behavior in future periods of time. The statistical methodologies presented in this article were applied on the parameters required for the weather generator. For the weather generation the following parameters must be provided for the 12 months of the year: averages of maximum daily temperature and minimum daily temperature, average of precipitation, maximum rain intensity in 30 minutes, number of rainy days, probability of dry day after wet day, probability of wet day after dry day, average solar radiation, and average dew point. And standard deviations of maximum daily temperature, minimum daily temperature, and precipitation, also skew coefficient of precipitation is required.

For the present project instead of using individual monthly values, monthly averages and monthly standard deviations for the four seasons were calculated for every parameter.

This way an initial data matrix of 990 rows (one row per weather station) by 39 columns was used for the cluster analysis.

Generalities of cluster analysis.

Cluster analysis is a multivariate statistical methodology used with the purpose of grouping individual sample points in groups called clusters. The methodology is based in similarity measurements between the sample elements employed in the analysis. The objective of clustering techniques is to maximize within group similarity while minimizing between group similarity. There exist many cluster methodologies, they differ considerably in their basic principles and in the results that they yield. Some clustering methods are divisive and dissect large data sets into smaller groups, others are agglomerative that start with individual data points that eventually are put together in several groups. Additionally, cluster methods may either be hierarchical, when a group at one level is combined with another group at a higher level (at the end of the procedure a large group is formed containing all the observations in the sample), or nucleated, when data from the sample is strongly differentiated and distinct cluster are very apparent. (Kalkstein, et. al., 1987).

Considering that data from weather stations all over the US is not strongly nucleated and that agglomerative procedures are more efficient with large sample sets, only the three most commonly used hierarchical and agglomerative clustering methods were initially tried for the climatic input data of the simulation model. These methods are Ward's, average linkage, and centroid. In initial tests using the three procedures, average linkage, and centroid yielded many clusters containing only one weather station, Ward's yielded two clusters containing only one weather station each.. One cluster contained only the weather station from Mount Washington in Maine and the other contained the weather station from the Death Valley in California. The number of weather stations in the other clusters produced by Ward's was very variable. The capability of Ward's method to isolate weather stations with extreme climatic characteristics, and the variable size of clusters were taken as evidences of Ward's good performance for clustering weather stations in this project. The two one-station clusters from Ward's were later discarded for further analysis because they are not locations of agricultural interest.

The similarity used in Ward's is the within cluster sum of squares summed over all the variables (W). It is presented by Kalkstein et al (1987) as follows:

$$W = \sum_{k=1}^K \sum_{j=1}^J \sum_{i=1}^{N_k} (X_{ijk} - X_{\cdot jk})^2 \quad (1)$$

Where X_{ijk} is the value of the j^{th} variable for the i^{th} of N_k observations in the k^{th} of K clusters. The dot indicates the mean value for that subscript.

W can be decomposed as:

$$W = \underbrace{\sum_{k=1}^K \sum_{j=1}^J \sum_{i=1}^{N_k} (X_{ijk} - X_{.j.})^2}_A - \underbrace{\sum_{k=1}^K N_k \sum_{j=1}^J (X_{.j.} - X_{.jk})^2}_B \quad (2)$$

The original sum of squares in equation (1) is composed of two parts: A, represents a new sum of squares with respect to a "new" centroid $X_{.j.}$, located somewhere between the original centroid (X_{jk}) of the two clusters under consideration for merging. And B, the squared Euclidean distance between the original centroids for the two pre-merger clusters and the new centroid weighted by N_k (the number of observations). Ward's method assumes that if two clusters 1 and 2 merge, the change in W (ΔW) after merging will depend only in the relationship between these two clusters and not on relationships between any other clusters.

$$\Delta W = W' - W \quad (3)$$

Where W is similarity before merge and W' is similarity after merge. Ward's method merges clusters with the smallest ΔW .

Cluster Analysis with SAS

Statistical analysis were performed with SAS Windows 6.11, and the SAS documentation consulted were the SAS/STAT User's Guide, release 6.03, and the SAS User's Guide: Basics, Version 5 Edition.

PROC CLUSTER was the procedure used for grouping weather stations. The statements used are as follows:

```
PROC CLUSTER DATA=LAST0 NOPRINT STD METHOD=WARD CCC PSEUDO
    OUTTREE=CLUS1;
VAR PRECIP TEMP SOLRA DEWP;
COPY STATE WSID WSNAME;
```

Options specific of the PROC CLUSTER statement are: STD, to standardize the original data to mean 0 and standard deviation 1 before starting the clustering procedure. METHOD=, to specify the clustering method. CCC, to print the cubic clustering criterion and approximated R^2 . PSEUDO, to print pseudo F and t^2 . OUTTREE= To specify a SAS

file name that will contain results of the cluster procedure to be used in the construction of a tree diagram. The statistics obtained with the CCC and PSEUDO options are used to evaluate the performance of the cluster procedure.

Words following the VAR statement are macro names. Each macro contains the seasonal variables associated with the macro name.

The COPY statement is followed by the variables from the input data sets that are to be included in the OUTTREE data set, those variables are state code, weather station numerical id, and the weather station name.

After the PRO CLUSTER procedure, the PROC TREE procedure is used to obtain the hierarchical structure resulting from the previous clustering procedure. The PROC TREE procedure was used with the following statements, and options.

```
PROC TREE DATA=CLUS1 NOPRINT OUT=CL1 NCL=50;  
COPY STATE WSID WSNAM PRECIP TEMP;
```

The CLUS1 data set specified in the input data set for the PROC TREE procedure is the output data set from the PROC CLUSTER procedure. The NOPRINT option was used to avoid printing the tree diagram, because it is very large and occupies several pages which makes very difficult to observe the hierarchical structure of the clusters. The OUT= option gives the name of a SAS file containing the results of the procedure. NCL= specifies the number of clusters to be included in the output data set.

To evaluate the performance of the cluster analysis the output data set from PROC TREE was sorted by cluster, state and weather station id using PROC SORT, and then it was printed with PROC PRINT. The print out was used to verify cluster membership of the weather stations. Quality of the clusters was also evaluated with maps depicting the clusters obtained in the cluster analysis. Clusters were judged as appropriate or inappropriate intuitively, based in geographical location, form, orientation, and weather station composition of clusters represented in maps. Maps were obtained using the Geographical Information System (GIS) named Graphical Resources Analysis Support System, GRASS, (US Army, 1988).

The set of procedures from PROC CLUSTER to PROC PRINT and the creation of a map were repeated several times modifying the input variables for the PROC CLUSTER, and the number of clusters specified in the NCL option of PROC TREE. The adequate number of clusters for each set of variables tested was decided based in graphs of R^2 against number of clusters, and CCC vs. number of clusters using PROC PLOT. A copy of the entire SAS program for climate grouping is shown in Appendix A.

After selecting the best weather station clusters, an hypothetical weather station will be created with the mean vector (set of monthly means from all the climate characteristics

used for the weather generation) of every cluster. EPIC will be run for each cluster using data from the hypothetical stations as the climate input.

SOIL INPUT

EPIC requires numerous inputs to describe the soil and its behavior. Two types of information must be furnished. First, up to eleven variables provide general information about the soil to be used in the simulation. Second, up to twenty variables are used to describe the physical and chemical characteristics of each identifiable layer (up to ten) in the profile (US Department of Agriculture, and Texas Agricultural Experiment Station, 1993).

The general soil characteristics used in this study were slope, and soil albedo. The following specific properties for the two top layers: Thickness, moist bulk density, water content at wilting point, water content at field capacity, sand content, silt content, pH, organic carbon content, calcium carbonate content, cation exchange capacity, rock content, dry bulk density, and hydraulic conductivity. And the following characteristics for the soil profile: Total water holding capacity, total organic carbon content, and total calcium carbonate content.. The statistical analysis started with a data matrix of 67, 291 rows (one row per soil) by 32 columns (one column for each soil property listed above).

The general steps followed to create soil groups that can be used for simulation models was:

- 1) Reduction of number of variables with Factor analysis.
- 2) Production of linear combinations between factors from factor analysis and original soil data.
- 3) Cluster analysis on the linear combinations data previously obtained.
- 4) Evaluation of soil clusters.

Generalities of Factor Analysis

The Factor Analysis is based on the following model that shows the original variables as linear combinations of several factors as shown in equation (4):

$$Y = X \beta + E \quad (4)$$

Where Y: matrix of original soil properties.

X: matrix of factor scores.

β : matrix of factor loads.

E: matrix of residuals.

The above model is similar to a multiple regression model expressed in matrix notation, with the difference that in this Factor Analysis model both β and X are unknowns. First

the factor loads must be found, and then solve the model for the factor scores (Srivastava, 1983) A major reason for doing Factor Analysis is to eliminate collinearity (redundancy) between original variables, and focus on only the variables with the greatest potential for explaining most of the total variance in the data. The loads are derived from the magnitude of eigenvalues associated with the correlation or covariance matrix of the original standardized variables. The solution of the model yields as many factors as initial soil properties, but only factors associated with eigenvalues larger or equal to one are used for further analysis.

FACTOR ANALYSIS WITH SAS.

Factor Analysis is performed by SAS with the PROC FACTOR procedure. Before running the Factor Analysis, data was standardized, with PROC STANDARD, to mean 0 and variance 1, with the purpose of removing any effect due to the different measurement units used for the soil characteristics. The statements and options used are as follows:

```
PROC STANDARD DATA=ONE OUT=STAN MEAN=0 STD=1 ;  
VAR SLOPE SALBDR MBD1 TIC1 SAN1 SIL1 CEC1 HC1 PH1 ROK1 DBD1 WP1  
FC1 OC1 CACO1 ROK2 SAN2 CEC2 HC2 SIL2 PH2 TIC2 MBD2  
DBD2 FC2 OC2 CACO2 WP2 TOC TCACO TWHC;
```

```
PROC FACTOR DATA=STAN OUTSTAT=FACT1 METHOD=PRINCIPAL  
SCORE;  
VAR SLOPE SALBDR MBD1 TIC1 SAN1 SIL1 CEC1 HC1 PH1 ROK1 DBD1 WP1  
FC1 OC1 CACO1 ROK2 SAN2 CEC2 HC2 SIL2 PH2 TIC2 MBD2  
DBD2 FC2 OC2 CACO2 WP2 TOC TCACO TWHC;
```

Data set STAN contains the standardized variables, and is the input data set in PROC FACTOR. The METHOD= option specifies the method for extracting the factors from the correlation matrix. The SCORE option allows to store the factor scoring coefficients in the output data set named FACT1.

The linear combinations between the factor scores and the original data are made with the PROC SCORE procedure as follows:

```
PROC SCORE DATA=STAN SCORE=FACT1 OUT=SCOR1;  
VAR SLOPE SALBDR MBD1 TIC1 SAN1 SIL1 CEC1 HC1 PH1 ROK1 DBD1 WP1  
FC1 OC1 CACO1 ROK2 SAN2 CEC2 HC2 SIL2 PH2 TIC2 MBD2  
DBD2 FC2 OC2 CACO2 WP2 TOC TCACO TWHC;
```

The DATA= and SCORE= options indicate the data sets that contain the standardized soil variables, and the factor scores respectively for calculation of the linear combinations between the two types of variables. The linear combinations are stored in the data file named SCORE1, the variables resulted from the linear combinations are named

FACTOR1, FACTOR2, FACTOR3....etc., depending in the number of factors selected from PROC FACTOR.

SOIL CLUSTERING WITH SAS.

The data matrix employed for soil grouping is the output from the PROC SCORE procedure, this is a matrix of 67,291 rows by 10 columns. For large data sets like the one used in this study PROC FASTCLUS is more efficient than PROC CLUSTER. This procedure uses a method that Anderberg (1973) called 'nearest centroid sorting'. A set of points called cluster seeds is selected as a first guess of the means of the clusters. Each observation is assigned to the nearest seed to form temporary clusters. The seeds are then replaced by the means of the temporary clusters, and the process is repeated until no further changes occur in the cluster (SAS Institute Inc., 1988).

Syntax and some of the options of the cluster procedure follows:

```
PROC FASTCLUS NOPRINT DATA=SCOR1 OUT=CLUS1 IMPUTE MAXC=&N;  
VAR FACTOR1--FACTOR10;  
ID SEC;
```

The output data set named CLUS1 will have the input variables, cluster numbers and several statistics useful for the evaluation of the cluster performance. The over-all R^2 is the statistic used with that purpose in this study. The option MAXC= is used to indicate the number of disjoint clusters that will be produced by the procedure. The symbol &N after MAXC is a macro variable, that change values to test the clustering procedure with different number of clusters. The option IMPUTE replaces missing values by the cluster average. The statement VAR is followed by input variables that are the linear combinations yielded by PROC SCORE. The statement ID is followed by SEC the soil sequence number which is one of the classification variables for identification of the soils.

Over-all R^2 obtained from consecutive runs of PROC FASTCLUS with increasing number of clusters was used to decide about the number of clusters to group soils for the EPIC input. Complete program is shown in Appendix B.

RESULTS

After careful examination of geographical location, and composition of climate clusters on maps it was decided that the best grouping of weather stations to be used as input of the model EPIC, was reached using only precipitation means and temperature means. Other variables such as solar radiation, dew point, and skew coefficient of precipitation are highly correlated with precipitation and temperatures and were causing "confusion" in cluster's construction, specially due to multicollinearity.

Appropriateness, of the cluster construction was achieved when weather stations were well grouped, and clusters appeared in the right position on the maps. This evaluation was done employing several reference locations. One of those locations was south Texas. Initially weather stations of south west Texas which have low precipitation were assigned to the same cluster with weather stations located in the gulf coast of Texas, which has considerably higher precipitation. That was the case, for example, of weather stations from Del Rio in south west Texas and Corpus Christi in the gulf coast. Changes in climatic variables and number of clusters were introduced until south Texas was separated in two different clusters. Similar scrutiny and variations in the analysis were done in other reference regions in the US.

After deciding about the best climatic variables for the clustering, then several cluster numbers were tried. Results of the tests are presented in Figure 1, and Figure 2. In Figure 1, R^2 does not have significant increase after 50 clusters, and in Figure 2, CCC presents the largest pick in 46 clusters. Considering results from R^2 and CCC criteria, it was decided that 50 clusters is an appropriate number of cluster to group weather station, explaining approximately 92% of the total variability contained in the data. Results of Climate clusters is illustrated in Figure 3, for the east part of the US.

Cluster procedure for soil grouping was based in ten factors produced by the Factor Analysis with eigenvalues equal or larger to one. An adequate number of cluster was obtained running the clustering program for several number of clusters. Results of this procedure are presented in Figure 4. Increasing the number of clusters beyond 100 do not increase values of Over-all R^2 . A significant increase of Over-all R^2 eventually would be achieved with a large number of clusters, 300 or more, which would be impractical for the purposes of constructing an input data set for EPIC. Based in Figure 4 it was decided that 100 clusters, explaining 70% of the total variability contained in the soil data are appropriate for construction of a soil input data set for EPIC. Maps depicting soil clusters will be constructed, and overlaid with the climate cluster map to obtain the climate-soil EPIC input.

Figure 1. R-square plot to decide about the number of clusters.

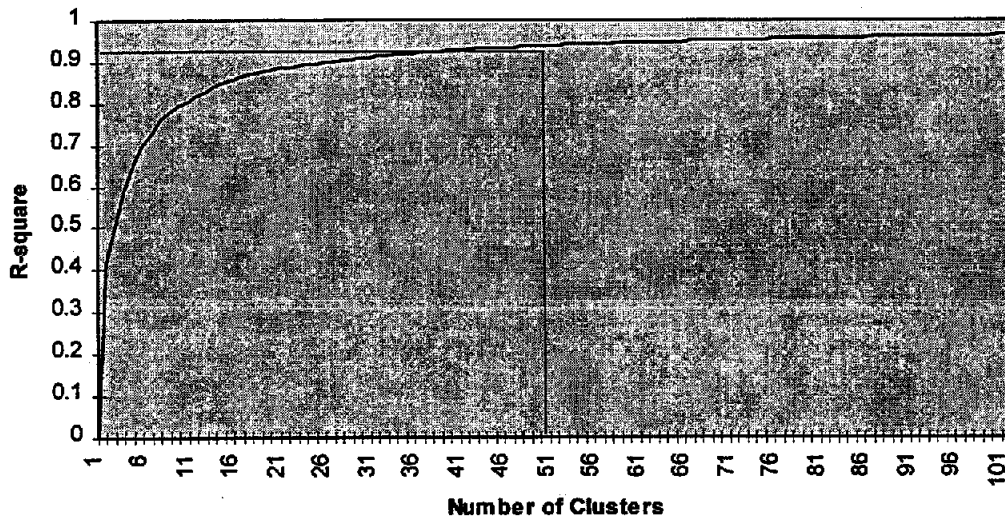
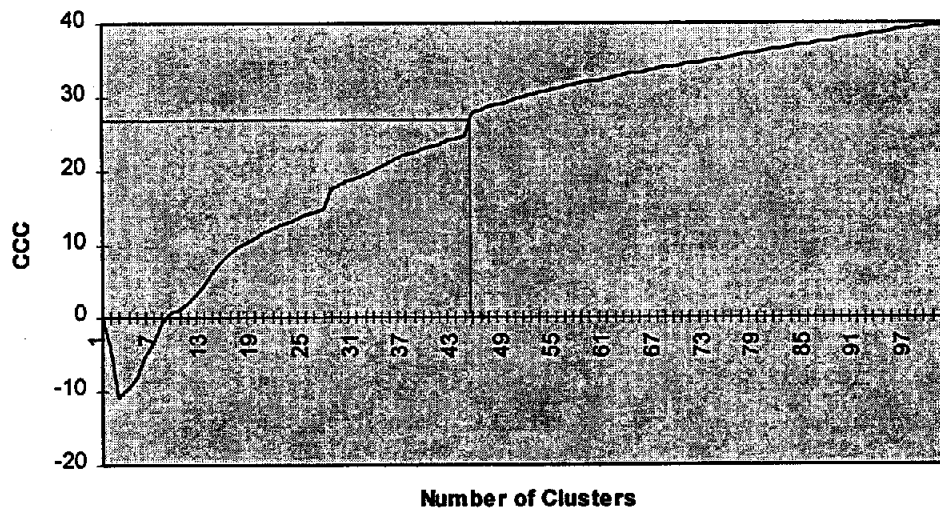


Figure 2. Cubic Cluster Criterion (CCC) curve to decide about cluster number.



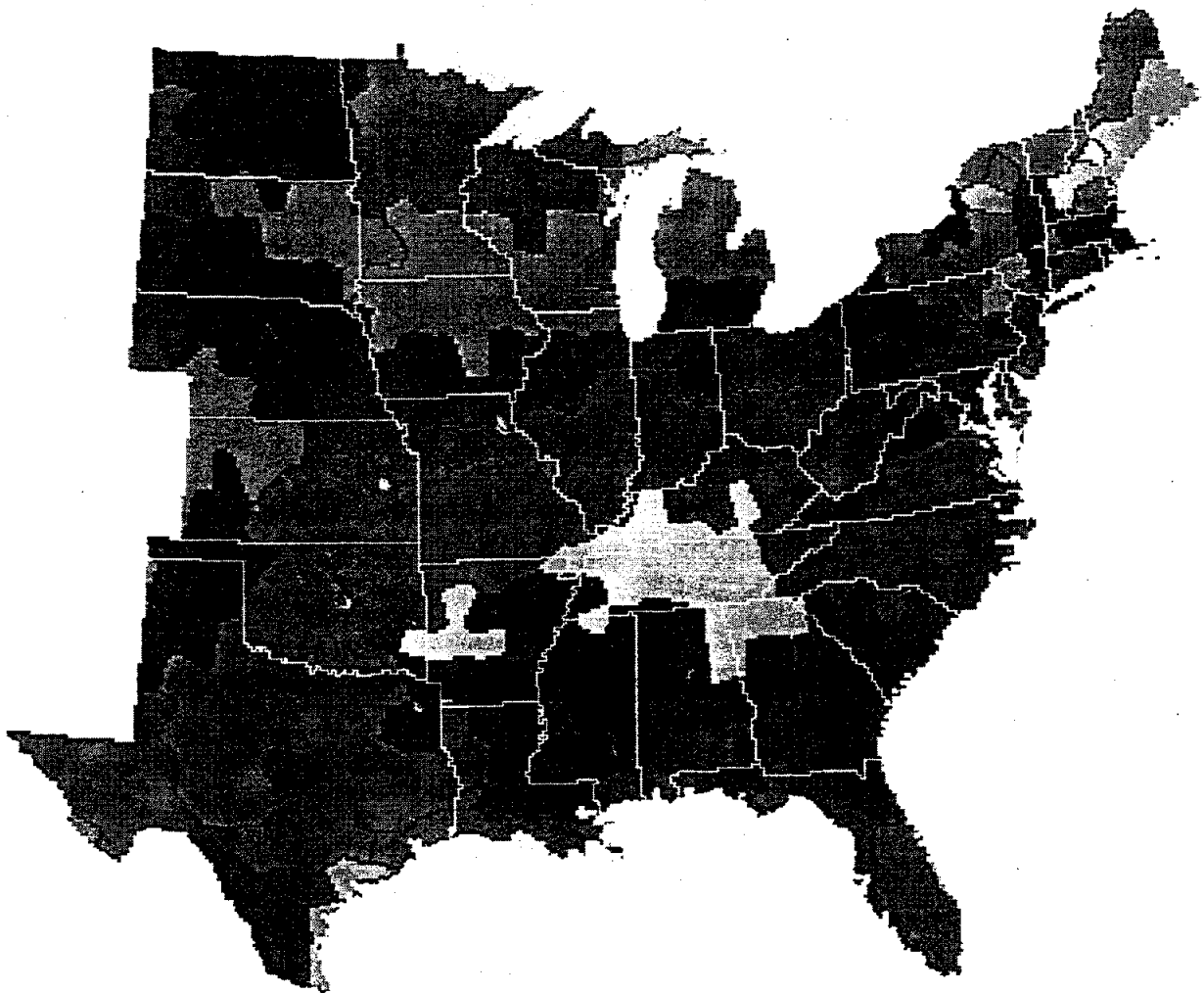
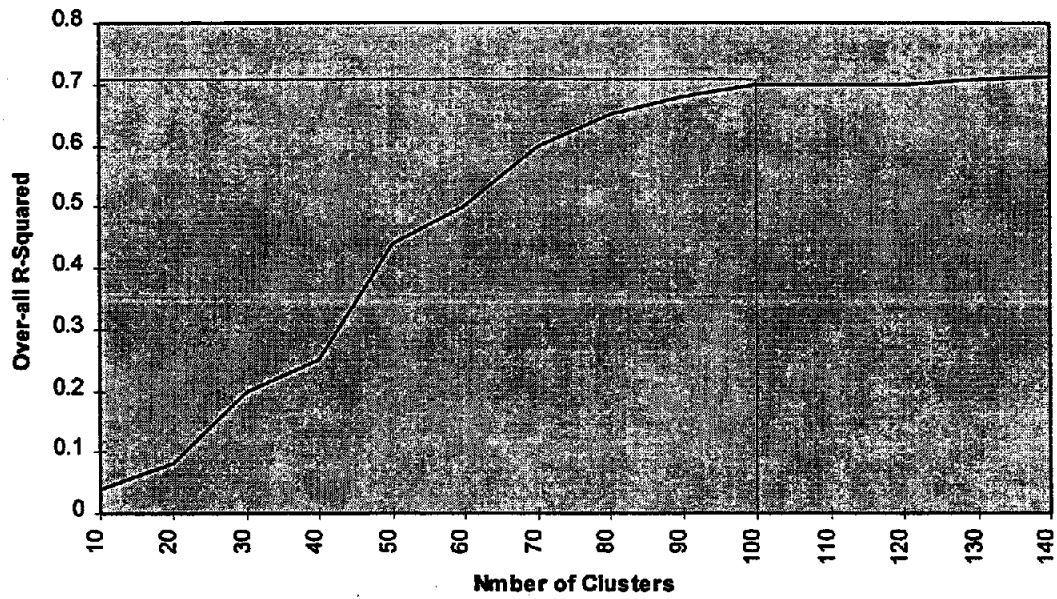


Figure 3 Weather Station Clusters for EPIC Input

Figure 4. Over-all R Squared plot to decide about the number of soil clusters.



REFERENCES

1. Anderberg, M.R. (1973). Cluster Analysis for Applications, Academic Press Inc., New York. In: SAS Institute. 1988. SAS/STAT User's Guide. Release 6.03 Edition. SAS Inst. Cary, NC.
2. Arnold, J.G., P. Allen, G. Bernhardt. (1993). A comprehensive surface-groundwater flow model. *Journal of Hydrology*, 142 p 47-69.
3. Kalkstein, L.S., G. Tan, and J.A. Skindlov. (1987). An Evaluation of Three Clustering Procedures for Use in Synoptic Climatological Classification. *Journal of Climate and Applied Meteorology*. 26 p. 717-730.
4. SAS Institute. 1988. SAS/STAT User's Guide. Release 6.03 Edition. SAS Inst. Cary, NC.
5. Srivastava, M.S. 1983. An Introduction to Applied Multivariate Statistics. North Holland. New York, NY. Pp 85-115.
6. U.S. Department of Agriculture, and Texas Agricultural Experiment Station. (1993). EPIC user's Guide.
7. U.S. Army. (1987). GRASS reference manual. USA CERL, Champaign, IL.
8. Williams, J.R., P.T. Dyke, W.W. Fuchs, V.W. Benson, O.W. Rice and E.D. Taylor. EPIC-Erosion Productivity Impact Calculator: 2. User Manual. USDA Technical Bulletin N° 1768. Washington D.C.: Agricultural Research Service, U.S. Department of Agriculture. 1990.
9. Williams, J.R., J. G. Arnold, R. Srinivasan, and T. S. Ramanarayan. APEX: A new tool for predicting the effect of climate and CO₂ changes on erosion and water quality. In: Boarman, J. and D.T. Favis-Mortlock (Eds.). *Modeling Soil Erosion*. Springer-Verlag. NATO-ASI Global Change Series. 1996.

APPENDIX A. SAS PROGRAM FOR CLIMATE CLUSTERS.

OPTIONS LS=80 PS=50;

MACRO PRECIP MPREWN MPRESP MPRESU MPREFL %
MACRO TEMP MMXTWN MMXTSP MMXTSU MMXTFL MMNTWN MMNTSP
MMNTSU MMNTFL %
MACRO SOLRA MSRWN MSRSP MSRSU MSRFL %
MACRO DEWP MDPWN MDPSP MDPSU MDPFL %

DATA USA;INFILE 'C:\NPLOSS\CLIMUSA1.TXT';
INPUT STATE \$ 1-2 WSID 3-5 WSNAME \$ 7-25 CHAR \$ 27-33 CHARID 35-36 +1 M1 M2
M3 M4 M5 M6 M7 M8 M9 M10 M11 M12;

IF WSNAME='MOUNT WASHINGTON WB' THEN DELETE;
IF WSNAME='DEATH VALLEY' THEN DELETE;
TOTYEAR=SUM(OF M1-M12);
AVGYEAR=TOTYEAR/12;

%MACRO SEASON(N,VAR);

DATA USA2;SET USA;

IF CHARID=&N THEN DO;
T&VAR.WN=(M1+M2+M12);
T&VAR.SP=(M3+M4+M5);
T&VAR.SU=(M6+M7+M8);
T&VAR.FL=(M9+M10+M11);
&VAR.WN=T&VAR.WN/3;
&VAR.SP=T&VAR.SP/3;
&VAR.SU=T&VAR.SU/3;
&VAR.FL=T&VAR.FL/3;
END;
PROC SORT;BY STATE WSID;

PROC TRANSPOSE DATA=USA2 OUT=TRANS1 ;BY STATE WSID WSNAME ;
VAR &VAR.WN ;
ID CHAR;

PROC TRANSPOSE DATA=USA2 OUT=TRANS2 ;BY STATE WSID WSNAME ;
VAR &VAR.SP ;
ID CHAR;

PROC TRANSPOSE DATA=USA2 OUT=TRANS3 ;BY STATE WSID WSNAME ;
VAR &VAR.SU ;
ID CHAR;

PROC TRANSPOSE DATA=USA2 OUT=TRANS4 ;BY STATE WSID WSNAME ;

VAR &VAR.FL ;
ID CHAR;

DATA ONE ;SET TRANS1;
&VAR.WN=&VAR;

DATA TWO ;SET TRANS2;
&VAR.SP=&VAR;

DATA THREE;SET TRANS3;
&VAR.SU=&VAR;

DATA FOUR ;SET TRANS4;
&VAR.FL=&VAR;

DATA UNION&N;MERGE ONE TWO THREE FOUR;BY STATE WSID WSNAME ;
DROP &VAR;

PROC DATASETS;
DELETE TRANS1 TRANS2 TRANS3 TRANS4 ONE TWO THREE FOUR;
RUN;
%MEND SEASON;

%SEASON(1,MMXT)
%SEASON(2,MMNT)
%SEASON(3,SMXT)
%SEASON(4,SMNT)
%SEASON(5,MPRE)
%SEASON(6,SPRE)
%SEASON(7,SKPR)
%SEASON(8,MPWD)
%SEASON(9,MPWW)
%SEASON(10,MNRD)
%SEASON(11,MI_5)
%SEASON(12,MSR)
%SEASON(13,MDP)

DATA FINAL;
MERGE UNION1 UNION2 UNION3 UNION4 UNION5 UNION6 UNION7 UNION8 UNION9
UNION10 UNION11 UNION12 UNION13;
BY STATE WSID WSNAME;

DROP MMXT MMNT SMXT SMNT MPRE SPRE SKPR MPWD MPWW MNRD MI_5 MSR
MDP;

PROC SORT;BY STATE WSNAME;

DATA COOR;INFILE 'C:\NPLOSS\COORUSA.TXT';

```

INPUT STATE $ 1-2 WSNAME $ 4-22 +1 LON +1 LAT;

PROC SORT;BY STATE WSNAME;

DATA LAST0;MERGE FINAL COOR;BY STATE WSNAME;
IF STATE='AP' THEN DELETE;
IF WSNAME='1AURORA RESEARCH FM' THEN DELETE;
IF WSID=381 THEN DO; LON=76.65;
    LAT=42.73;END;

IF WSNAME='LAKE CHARLES WSO' THEN DO; LON=93.22;
    LAT=30.12;END;

PROC DATASETS;
DELETE UNION1-UNION13;

PROC CLUSTER DATA=LAST0 NOPRINT STD OUTTREE=CLUS1 METHOD=WARD
CCC PSEUDO ;
VAR PRECIP TEMP ;
COPY STATE WSID WSNAME;

PROC TREE NOPRINT DATA=CLUS1 PAGE=1 OUT=CL1 HEIGHT=N MAXH=200 HOR
SORT NCL=50;
COPY STATE WSID WSNAME PRECIP TEMP ;

PROC SORT DATA=CL1;BY STATE WSID CLUSTER;

DATA FINAL;SET CL1;
SEP1='=';

FILE 'F:\NPLOSS\CLUSCL1.TXT' LS=500;
PUT STATE $ SEP1 $ WSNAME $ SEP1 $ WSID SEP1 $ CLUSTER ;

RUN;

```

APPENDIX B. SAS PROGRAM FOR SOIL CLUSTERS

```
OPTIONS LS=80 PS=60;
```

```
PROC DATASETS KILL;
```

```
%MACRO CLUS(N);
```

```
DATA ONE;INFILE "D:\NPLOSS\SOILDATA\SOILUSA1.TXT" FIRSTOBS=2 DLM="|"  
MISSOVER;
```

```
INPUT SEC STATE $ COUNTY NUMS5 $ SLID $ SLOPE HYDRGP $ SALBDR TIC1 MBD1  
WP1 FC1 SAN1 SIL1 PH1 OC1 CACO1 CEC1 ROK1 DBD1 HC1 LAYER2 TIC2 MBD2 WP2  
FC2 SAN2 SIL2 PH2 OC2 CACO2 CEC2 ROK2 DBD2 HC2 TWHC TOC TCACO;
```

```
PROC SORT;BY SEC;
```

```
PROC PRINT;
```

```
PROC STANDARD DATA=ONE OUT=STAN MEAN=0 STD=1 ;
```

```
VAR HC2 HC1 ROK2 SAN2 SAN1 CEC2 SIL1 SIL2 CEC1 ROK1 SLOPE PH2 PH1 TIC2  
MBD2 MBD1 TIC1 DBD2 DBD1 FC2 SALBDR TCACO FC1 TWHC CACO2 WP2 WP1 TOC  
CACO1 OC1 OC2;
```

```
PROC FACTOR DATA=STAN OUTSTAT=FACT1 METHOD=PRINCIPAL SCORE;
```

```
VAR HC2 HC1 ROK2 SAN2 SAN1 CEC2 SIL1 SIL2 CEC1 ROK1 SLOPE PH2 PH1 TIC2  
MBD2 MBD1 TIC1 DBD2 DBD1 FC2 SALBDR TCACO FC1 TWHC CACO2 WP2 WP1 TOC  
CACO1 OC1 OC2;
```

```
PROC SCORE DATA=STAN SCORE=FACT1 OUT=SCOR1;
```

```
VAR HC2 HC1 ROK2 SAN2 SAN1 CEC2 SIL1 SIL2 CEC1 ROK1 SLOPE PH2 PH1 TIC2  
MBD2 MBD1 TIC1 DBD2 DBD1 FC2 SALBDR TCACO FC1 TWHC CACO2 WP2 WP1 TOC  
CACO1 OC1 OC2;
```

```
PROC FASTCLUS NOPRINT DATA=SCOR1 OUT=CLUS1 IMPUTE SHORT MAXC=&N;
```

```
VAR FACTOR1--FACTOR10;
```

```
ID SEC;
```

```
DATA LAST;SET CLUS1;
```

```
PROC SORT;BY SEC;
```

```
DATA FINAL;MERGE ONE LAST;BY SEC;
```

```
DEL="|";
```

```
FILE "F:\NPLOSS\CLSOIL.&N..DAT" ;
```

```
PUT SEC DEL $ STATE $ DEL $ COUNTY DEL $ NUMS5 $ DEL $ SLID $ DEL $ LUSTER;
```

```
PROC DATASETS KILL;
```

```
%MEND CLUS;
```

```
%CLUS(10)
```

```
%CLUS(20)
```

```
%CLUS(30)
```

%CLUS(40)
%CLUS(50)
%CLUS(60)
%CLUS(70)
%CLUS(80)
%CLUS(90)
%CLUS(100)
%CLUS(110)
%CLUS(120)
%CLUS(130)
%CLUS(140)