1. Publication Nº	2. Version	3. Date	5. Distribution
INPE-2490-PRE/173		July, 1982	☐ Internal ☒ External ☐ Restricted

4. Origin	Program
DSE/DES	INTEG

6. Key words - selected by the author(s)

CLUSTER ANALYSIS
HOMOGENEOUS SUBREGIONS

7. U.D.C.: *528.711.7:551.5:63*

8. Title *INPE-2490-PRE/173*

AN APPLICATION OF CLUSTER ANALYSIS FOR DETERMINING HOMOGENEOUS SUBREGIONS: THE AGROCLIMATOLOGICAL POINT OF VIEW

10. Nº of pages: *10*

11. Last page: *09*

12. Revised by

Getulio Teixeira Batista

9. Authorship *Carlos Alberto Cappelletti*

Responsible author

13. Authorized by

Nelson de Jesus Parada
Director

14. Abstract/Notes

A stratification oriented to crop area and yield estimation problems was performed using an algorithm of clustering. The variables used were a set of agroclimatological characteristics measured in each one of the 232 municipalities of the State of Rio Grande do Sul, Brazil. A non-hierarchical cluster analysis was used and the pseudo F-statistics criterion was implemented for determining the "cut point" in the number of strata.

15. Remarks *This paper was accepted for presentation at the Sixteenth International Symposium on Remote Sensing of Environment, Buenos Aires, Argentina, June 2-9, 1982.*

AN APPLICATION OF CLUSTER ANALYSIS FOR DETERMINING HOMOGENEOUS SUBREGIONS: THE AGROCLIMATOLOGICAL POINT OF VIEW*

Carlos Alberto Cappelletti

Instituto de Pesquisas Espaciais - INPE
Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq
Caixa Postal 515, 12200 - São José dos Campos, SP, Brazil

ABSTRACT

A stratification oriented to crop area and yield estimation problems was performed using an algorithm of clustering. The variables used were a set of agroclimatological characteristics measured in each one of the 232 municipalities of the State of Rio Grande do Sul, Brazil. A nonhierarchical cluster analysis was used and the pseudo F-statistics criterion was implemented for determining the "cut point" in the number of strata.

1. INTRODUCTION

In order to predict the crop production of a region it is necessary to estimate two parameters: CA and P, crop area and yield, respectively, and integrate them by the expression

$$TP = CA * P \qquad (1)$$

where TP indicates total production.

The estimation of CA and P depends on a data set which can be obtained by different means: statistical sampling system or a census data collecting system in the area of interest. Aerial photograph and/or LANDSAT imagery are important means for CA estimation; for P, a yield prediction model can be implemented, see Baier (1979) and Cappelletti et al. (1981).

Both of the above approaches requires the stratification of the area to be studied for producing an adequate confidence coefficient in the final estimates (Raj, 1968).

This paper reports results obtained in the construction of strata with the application of an algorithm of Cluster Analysis to a set of data consisting of agroclimatological variables, whose values are, in general, averages values of historical series.

The data refers to each one of the 232 municipalities of the State of Rio Grande do Sul, Brazil. The decision to consider as a unity such political division was because the data were published in a municipality level.

The building of homogeneous strata is the first step in a project of crop production estimation. Since the parameters AC and P in Equation (1) depends on different characteristics, two set of strata are required.

---

*Presented at the Sixteenth International Symposium on Remote Sensing of Environment, Buenos Aires, Argentina, June 2-9, 1982.

2. METHODOLOGY

The technique of clustering has been widely used for grouping similar units. This technique works with a data matrix and a similarity measure.

The data matrix has dimension N * P, being N the number of units and P the number of characteristics observed or calculated for each unit.

The similarity measure used in this paper was the Euclidean distance in the observation space.

The objective of the clustering algorithm is to minimize the intracluster sum of squares following the K-means procedure of Mac Queen (1967).

The algorithm works as follows: the i^{th} unit of the j^{th} variable has value x(i,j), i = 1,...,N, j = 1,...,P, and each of the N unities lies in just one of K cluster. Denoting the mean of the j^{th} variable over the unities by $\bar{x}(\ell,j)$, the distance between the i^{th} unity and the s^{th} cluster is:

$$D(i,s) = \left[\sum_{j=1}^{P} (x(i,j) - \bar{x}(s,j))^2 \right]^{1/2}$$

and the error partition is

$$E\left[P_a(m,k)\right] = \sum_{i=1}^{N} D^2(i,s(i))$$

where s(i) is the cluster containing the i^{th} unity and P_a indicates a partition.

The procedure searchs for a partition with small E by moving unities from one cluster to another and ends when no such movement reduces E (Hartigan, 1975).

Clusters were generated in a nonhierarchical process. The process began with one group and stopped when the number of groups reached the cut point given by the pseudo F-statistics (PFS) criterion function of Vogel and Wong (1979).

The PFS criterion provided the optimal number of groups operating with the weighted ratio of the traces of the matrices B and W, respectively the matrices of sums of squares between and within groups in the multivariate analysis of variance (MANOVA) of the data set.

3. VARIABLES AND DATA SET

The variables used were:

wheat cultivated area (Ha)
wheat yield (kg/Ha)
average farm size (Ha)
agricultural land adequated for wheat (Ha)
average temperature (^{o}C)
potential evapotranspiration (mm)
rainfall in normal years (mm)
rainfall in dry years (mm)
average yearly run-off (%)
useful fraction of rainfall in normal years (mm)
useful fraction of rainfall in dry years (mm)
internal drainage in normal years (mm)
internal drainage in dry years (mm)

moisture deficit in normal years (mm)
moisture deficit in dry years (mm)
necessary minimum rainfall (mm)

The data set was taken from the Anuário Estatístico do Rio Grande do Sul (1968, 1969, 1970 and 1976) and from Ministério da Agricultura (1976).

4. RESULTS AND DISCUSSION

4.1 WHEAT AREA ESTIMATION (AC)

Two variables, relative crop area (RCA) and normalized average farm size (AFS), by municipality, were considered.

The reason for using those variables was that the first one represents the density of cultivated wheat lands and will provide homogeneous strata with respect to the importance of wheat in the agricultural scene. The average farm size was selected to represent problems which might be encountered in LANDSAT data classification with different field sizes.

The clustering algorithm gave four groups as the best partition (Table I).

STRATA No.	MEAN VALUES CRA(%)	AFS(Ha)	STANDARD DEVIATION CRA	AFS	No. OF MUNICIPALITIES
1	8.5	22.7	9.4	9.9	177
2	4.8	85.9	7.6	22.8	34
3	4.2	192.8	6.8	28.2	12
4	1.2	279.1	1.4	24.5	9

Table I. Four strata for CRA and AFS

The columns of the two mean values in Table I show that there exists a negative relationship between the variables CRA and AFS, that is, municipalities with large average farm size dedicate a large percentage of the total land for livestock-raising instead of wheat.

Figure 1 shows the geographical location of the strata.

In order to compare with results showed in Table I, the state was subdivided into four regions with variable CRA alone (Table II).

STRATA No.	MEAN	S.D.	No. OF MUNICIPALITIES
1	35.6	4.1	10
2	21.2	3.4	26
3	10.1	2.6	55
4	1.8	1.9	141

Table II. Four strata for variable CRA

Figure 2 shows the geographical location of these stratas.

4.2 WHEAT YIELD ESTIMATION (P)

The original data set used in this stratification included the last fourteen variables listed in item 3.

The data were those related with the wheat growing season and some of them were average values of historical series.

Those data were treated with a Principal Component Analysis (Cooley and Lohnes, 1971) and after that the scores factor for the first five principal components, which account for 93.5% of the total variance, feeded the clustering algorithm.

The technique of principal components have been described in the book cited. If should be recalled that the method produces uncorrelated linear functions of the original variables without any loss of information. Depending upon the data being used it may be possible to recognize the physical significance of these new variables. When this is possible, this method brings about a reduction in the quantity of basic data that needs to be used.

Table III shows the factor eigenvalues and the cumulated percent of the trace.

FACTOR	EIGENVALUE	CUMULATED PERCENT OF THE TRACE
1	7.69	54.9
2	2.05	69.6
3	1.73	81.9
4	1.01	89.2
5	0.60	93.5
OTHER 9 FACTORS		6.5

Table III. Output of the Principal Components Analysis. First five principal factors

With the factor score coefficients, which are an output of the principal components analysis, the factor scores for each municipality were calculated, and these data feeded the clustering algorithms.

Table IV shows the number of municipalities for each strata.

STRATA No.	No. OF MUNICIPALITIES
1	57
2	72
3	74
4	29

Table IV. Stratas for five principal components with fourteen agro-meteorological variables

Figure 3 shows the geographical position of the strata.

The more relevant characteristics for each strata can be summarized as follows:

Strata I.

This region is called "campanha" and corresponds to an extensive natural grassland zone traditionally used for livestock grazing.

Strata II.

This region is called "depressão Central" and corresponds to a small-scale diversified agriculture.

Strata III.

This region is called "planalto médio". It is the soybean-wheat cropping region in the state. The wheat growing season is the winter and this stratum corresponds to the largest yields of the state.

Strata IV.

In this region wheat is not cultivated because the urban and industrial zone of the great Porto Alegre, the state capital, is in it. Another region in this strata is the atlantic litoral where there are numerous areas of sand dunes. The others two subregions of this strata are in the livestock raising zone.

5. CONCLUSIONS

A stratification oriented to crop area and yield estimation problems was performed.

The algorithm of clustering used produced good results inasmuch as the geographic location of the strata appears to be logical and the strata seem to represent different conditions. Besides that, the within strata sum of squares was minimizes when a set of agro-meteorological variables was simultaneously considered.

The region chosen to apply the procedure has been extensively studied and consequently there exists the possibility of validating the criterium used. In order to improve the final results further work has to be done.

6. REFERENCES

Anuário Estatístico do Rio Grande do Sul, Secretaria de Economia, Departamento Estadual de Estatística, Vol. 1, 2 and 3, 1968, 1969 and 1970; Vol. 5.8 - tomo 1, Agropecuária, 1972-75, publicado em 1976, Porto Alegre.

Baier, W. Note on the terminology of crop weather models. Agricultural Meteorology, 20: 137-145, 1979.

Cappelletti, C.A.; Reis, J.R.; Lorena, L.A.N.; Dias, N.T.; Cruz Paião, L.B.F. da; Olivo, A.A. de; Costa, S.R.X. Proposta metodológica para a modelagem do crescimento de uma cultura visando estimação de produtividade agrícola. São José dos Campos, INPE, nov. 1981. (INPE-2255-PRE/037)

Cooley, W.W.; Lohnes, P.R. Multivariate data analysis. Wiley, New York, 1971. 364 p.

Hartigan, J.A. Clustering algorithms. Wiley, New York, 1975. 351 p.

Mac Queen, J.B. Some methods for classification and analysis of multivariate observations. Proceedings Symp. Math. Stat. and Prob., 5th, Berkeley 1, 281-297, 1967.

Ministério da Agricultura. Levantamento e avaliação de recursos naturais, sócio-econômicos e institucionais do Rio Grande do Sul. Vols. 1 to 6, INCRA, Brasília, 1976.

Raj, D. Sampling theory. McGraw-Hill, New York, 1968. 302 p.

Vogel, M.A.; Wong, A.K.C. PFS clustering method, IEEE Trans. on Pattern Anal. and Mach. Intel.. Vol. PAMI-1, No. 3, July 1979.

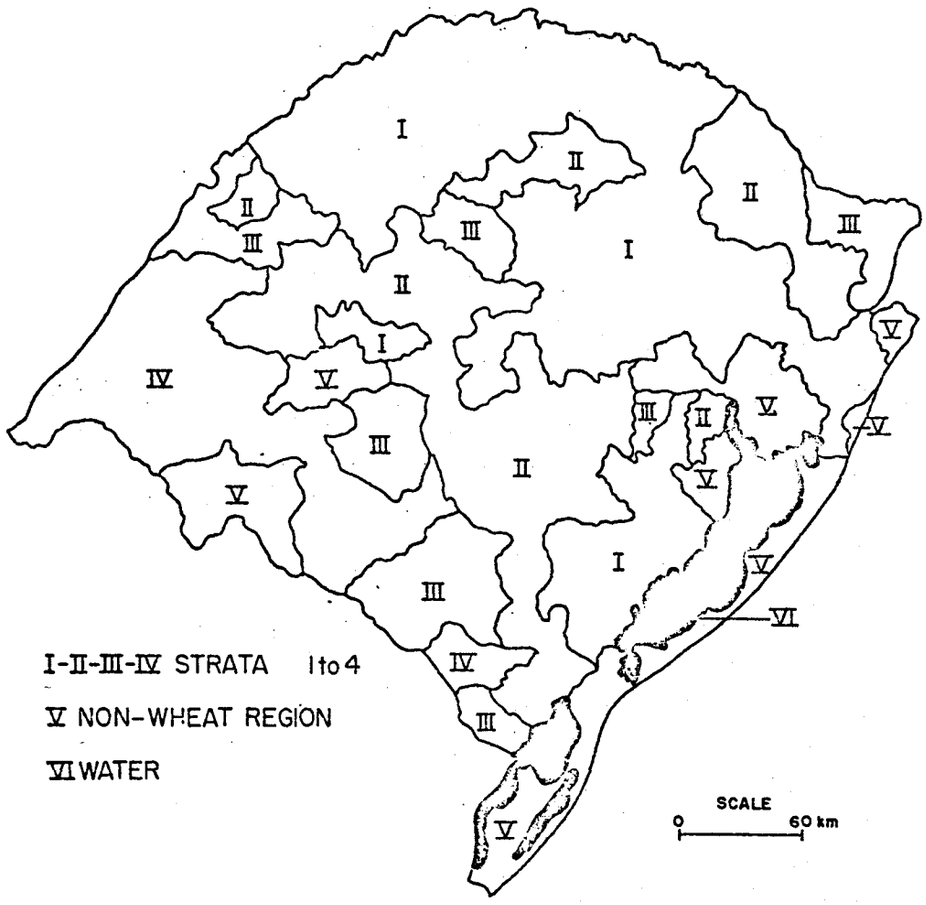

Figure 1. Four strata for CRA and AFS

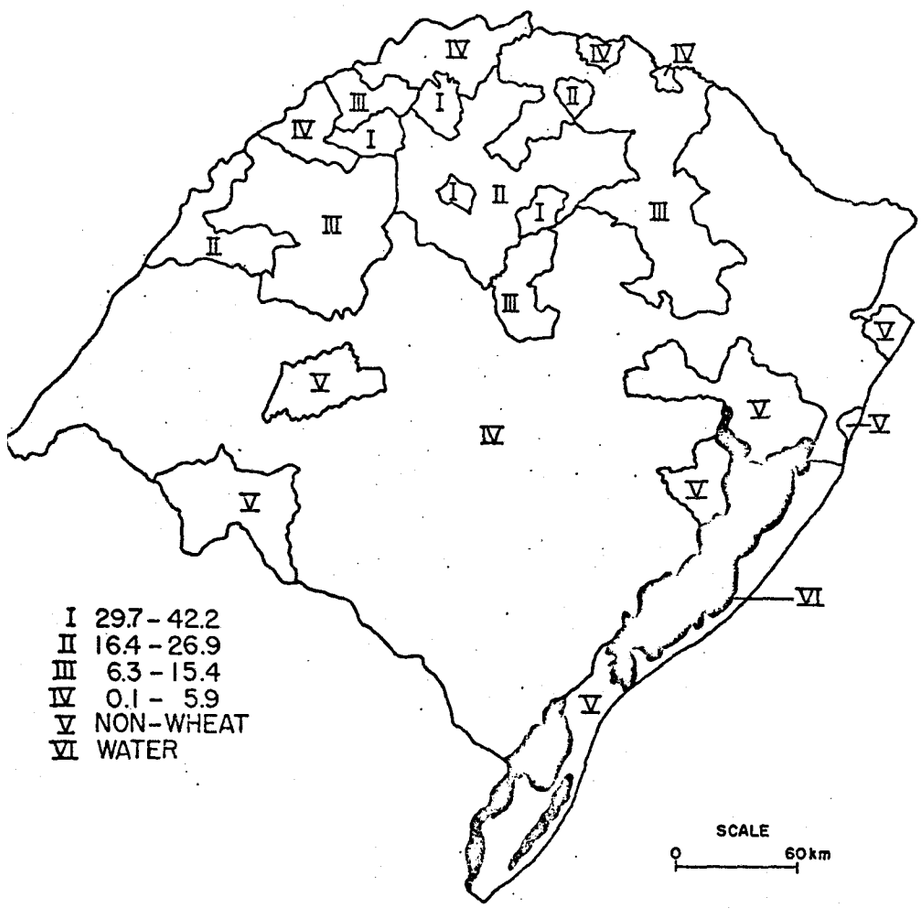

Figure 2. Four strata for CRA (%)

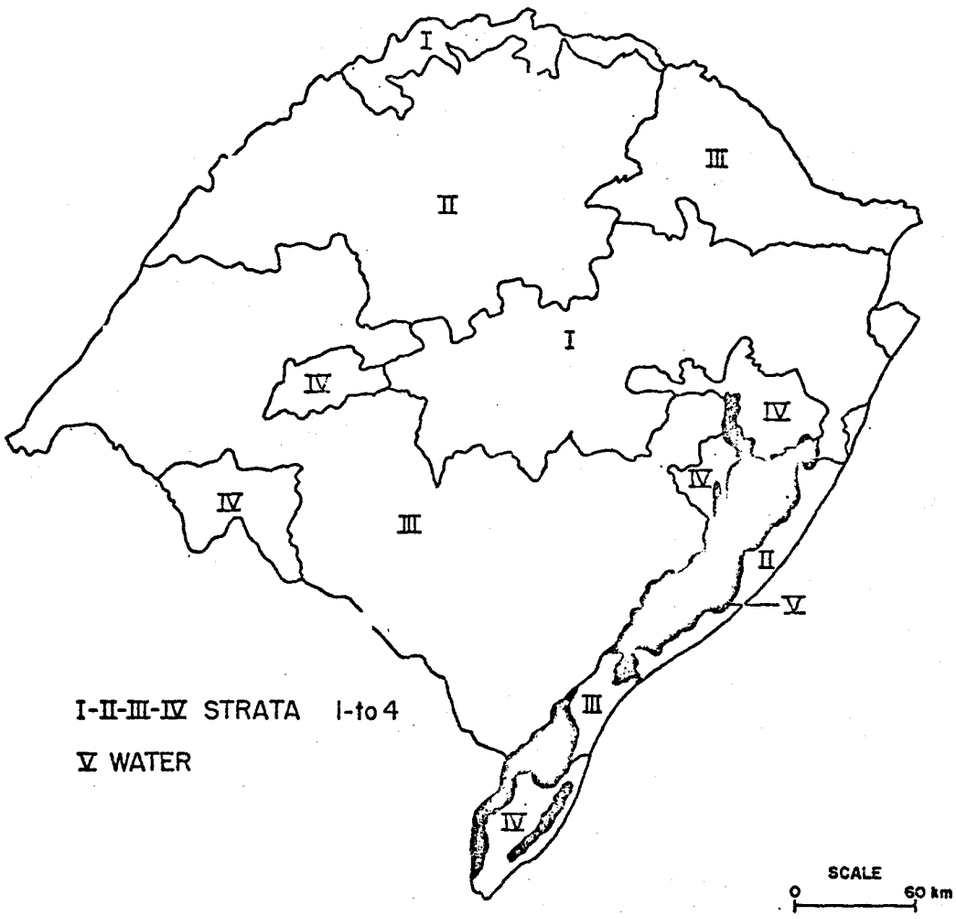

Figure 3. Four strata for five principal components with fourteen agro-meteorological variables