Aggregation in Input–Output Tables: How to Select the Best Cluster Linkage

B. Caber, D. Contreras, E.J. Miravete[†⋆]

April 12, 1991

Abstract

In this paper we try to give a solution to the aggregation problem on working with I–O Tables. First of all we verify the similarity degree among the production functions of industries which aggregate in each sector. Second, once we have established the aggregation by using different cluster analysis, we set a bundle of conditions to choose the proper linkage method. That allow us to characterize the way of aggregation (weighted or non–weighted) of the I–O Table. JEL: C67.

Keywords: Input–Output Tables, Aggregation, Linkage.

[†] Department of Economics, Universitat de València. P.O. Box 22006, València, Spain.

[⋆] The authors are very grateful to Professor Fontela for his useful comments.

1 Introduction

Aggregation in I-O Tables has to be done on the scope of the microeconomic foundations of macroeconomic analysis. This point is controversial and there are not convincing arguments which let us to keep a "golden rule." This is the logical consequence of dealing with so a difficult and complicated a subject; therefore the controversy is still open. The usual way to carry on aggregation is just to consolidate industries into sectors. There are some reasons that support that kind of aggregation based on the manageability of the data. The only justification we find to this point of view is the necessity of working with empirical data, but it has no theoretical interest or justification.

Related with the topic just mentioned there is another problem: the optimal size of an I–O Table, because depending on the degree of aggregation we can miss important information, on the other hand if the degree is quite low we can be violating homogeneity hypothesis. In Leontief's model, it is assumed the existence of productions functions with constant returns defined on an undetermined number of goods. These two characteristics are known as homogeneity and stability hypothesis and they are not completely compatible. This fact is due to the model aggregation degree: in this model the stability hypothesis is related with the fixed coefficients of the production functions of the industries, which ensure us the existence of constant returns. In order to keep the stability hypothesis we must aggregate industries into an adequate number of sectors (not too big), because if we have a table with a lot of industries each one producing only one homogeneous good, we can't assume, when the prices change, that all cross–elasticities are going to be zero, because with such a big amount of goods it has to be easy to find out substitutes among them. In order to keep stability hypothesis (i.e. fixed coefficients) we must aggregate industries, and this is the opposite effect required by the homogeneity hypothesis, because the bigger the aggregation degree is the greater is the set of heterogeneous goods encompassed in a sector. Hence it is almost impossible both hypothesis to be proved simultaneously.

Microfoundation problems are not treated in the above discussion, the only point it deals with is the coherence scope of a production function with fixed coefficients in an aggregated economy. Notwithstanding that problem wouldn't exist if industries would keep a fixed proportion among them in each sector, or the sectors grouping them would reflect their characteristics, that is the sectors would have the same production functions as the industries. Hatanaka (1952), Ara (1959) and Malinvaud (1954) papers studied that question setting the conditions to be accomplished by industries productions functions in order that, when they were aggregated into sectors, technic coefficients defined on sectors, were not affected for variations in the final demand vector. This is just known as Hatanaka's condition, and we express it as [Kossov (1972)]: $TA = A * T$, where A is the original coefficient matrix (on industries) and $A*$ is the aggregated coefficient matrix (on sectors). T is the aggregation matrix which can be unweighted (elements are 1 or 0) or weighted [Morimoto (1971)]. The economic meaning of this condition is absolutely clear. To aggregate industries it is necessary that they have homogeneous inputs structure, that is, a similar cost functions for each industry aggregated in a single sector. That is the

way to achieve the relationship between sectors (macroeconomic categories) and industries (microeconomic categories).

This analysis belongs to the class of exact aggregation models. The assumptions that this frame implies lead us to one of the two following points: either the aggregation is the correct one or it can't be done according to the designed T matrix. If Hatanaka's condition doesn't hold, the only solution is to redefine sectors in order to find a new aggregation matrix with no bias, that is, the technical coefficients of the aggregated model don't change when demand vector does. As it is a necessity to work with aggregated tables, methods of non-exact aggregation are developed. These methods don't need any requirements on the coefficient matrix but need a certain structure or behavior on the demand vector. We want to remark Theil's (1957) work who establish the conditions for zero first order bias, which implies to group industries into almost input-homogeneous sectors. There are other methods that try to minimize bias, [see a brief survey in Kymn and Norsworthy (1976)] but most of them, as those already mentioned, take the aggregation scheme as granted [Fisher (1962) and Neudecker (1970)]. The main feature of cluster analysis is to group industries into sectors depending on the degree of similarity on its cost functions characteristics. This method of aggregation generates endogenously the different sectors which industries group.

2 Cluster Analysis

Cluster analysis is used to classify data by clustering them, based on its similarities and/or disparities. These data are usually individuals, but in our case, we are dealing with industries' cost functions. The main difference of this framework in relation with the other multivariate analysis techniques is that the latter need previous grouping in order to classify individuals, while the former groups the data endogenously. Fisher's (1969) is one of the first to approach the problem. Now we try to find out the best grouping method: in order to get that, we apply different kinds of cluster analysis. We develop some simple conditions which allow us to verify the best method to aggregate using an autoregressive model on the distances found with the several clusters applied according to the linkage chosen. We apply these conditions to an Input–Output Table of the Spanish Economy.

The way the different clusters are set up depends on which similarity or disparity criteria are applied, on the algorithms we use and on the optimization procedure chosen. Putting together all the mentioned characteristics we can classify each kind of cluster.

It is necessary to use a measure of distance to quantify the divergence among individuals (cost functions in this paper). Well known distances are Euclidean, Euclidean normalized, Tchebycheff's metric, Mahalanobis', and Tonimoto's. So far as we are concerned, we choose Euclidean distance because it is easy to work it out and also the simplest. Particularly, it exhibits some very interesting properties for our goal; first, it is not invariant to scale change and second, it depends on the relationships among the variables. When

measuring the distances among clusters we consider the whole set of variables included in each cost functions. Once we have studied the distance among individuals we must work out the distance between group–individual (clustered sectors and industries) and among groups. To achieve that aim we have different criteria.

When doing a multivariate data analysis, we try to classify individuals by theirs characteristics. Cluster analysis allow us to group industries as if they were individuals to set different sectors. We use hierarchical cluster, that is, a subset of clustering techniques, because this particular kind of cluster let us to work without imposing any restriction about which industries are going to be in each cluster and besides to keep all the industries' characteristics. The most remarkable feature of hierarchical cluster is the fact that each group integrates only in one upper order cluster according to its scale dissimilarities; the greatest the degree of aggregation is, the greatest is the dissimilarity among groups. When that process ends up, there will be solely a cluster grouping all the individuals.

Each hierarchical cluster has a representation which is called *dendogram* that shows how clusters group (see Fig. 1) and also a similarity measure for each level. This sort of cluster works out the grouping in a very simple way because it starts forming so many clusters as individuals and choosing a similarity measure, as strategy, groups the most similar clusters. Once this first stage is accomplished it computes the gravity center of each new cluster and runs it again. Once we have established the technique and distance to use, we must choose what linkage method has to be applied, because the results are not independent of the linkage used. Linkage refers to how to measure distances among clusters' gravity center.

Depending on the linkage carried out, cluster analysis will have the *combinatority* and *compatibility* properties [Sneath and Sokal, 1973]. When we speak about *combinatority* and *compatibility* properties, we refer to the following conditions:

1. *Combinatority (C.1):* Let d^2_{ik} be the distance between clusters (i) and (k). If those are the ones which have the smallest distance, they achieve a new cluster (ik), with size $n_i + n_k$ (the summation of individuals in each group). The next step is to cluster a new group (ikh) with the previous clusters (ik) and (h), which exhibited the minimum distance (greater similarity). If $d^2_{(ik)h}$ can be computed just with d^2_{ik}, d^2_{ih}, $d^2_k h$, then this linkage posses the above mentioned property. In the particular case we are studying (Euclidean distance), this condition can be expressed as:

$$d^2_{(ik)h} = \alpha_i d^2_{ih} + \alpha_k d^2_{kh} + \beta d^2_{ik} + \tau \mid d^2_{ih} - d^2_{kh} \mid, \tag{1}$$

   where α_i, α_k, β and τ are the parameters to be found according to the linkage used.

2. *Compatibility (C.2):* This property holds when in equation (1) $\tau = 0$, and:

$$\alpha_i + \alpha_k + \beta = 1. \tag{2}$$

   This property means that distances among different patterns have the same characteristics, *i.e.*, distances became bigger in a uniform way according to the degree

of aggregation. This implies that all the distances have the same dimensionality, so they can be studied in a single model [see Escudero (1977)].

Depending on the parameters values we can compare heuristically different linkages when using Euclidean distance, as can be seen en Table 1. When using cluster analysis, different linkage methods are used but usually, reasons for choosing among them are not provided. Examining Fig. 1 we can check dendograms which cluster the same groups in very different ways according to the linkage chosen. These dendograms posses one or both of the mentioned properties. This characteristic is the one which makes the selection of the linkage to use not a trivial point, because in input–output tables there are industries absolutely different, so they can't be integrated in the same group at any cluster stage. In our analysis, the aggregation of industries in input–output tables, we want the cluster to satisfy just the property of combinatority but not the compatibility condition otherwise at every stage of the cluster, one or more industries would be clustered in the same group. This means that, even when the initial industries are different, they will be grouped in a single sector, hence the whole economy would be explained for the same model, *i.e.*, a unique cost function or production function. Working that way, a perfect homogeneous aggregation would be achieved, which is obviously not compatible with the multi–sectoral analysis of input–output tables.

Hence, we select linkage methods with economic sense. In our analysis we want the cluster to satisfy the above–mentioned properties (combinatority and no compatibility); there are three different linkages which have those properties: Centroid (unweighted pair group centroid), Median (weighted pair group centroid), and the Ward method [see Anderberg (1973)].

3 Testing

Since we have already set up the correct way to cluster, at a second stage we test the performance of the technique chosen. To carry out that test we estimate an autoregressive model with constant coefficients. The model we estimate is based in equation (1), which is referred to as an $N-1$ level of aggregation. The same combinatority condition for level N can be expressed as:

$$d^2_{(ikh)j} = \alpha_{ik} d^2(ik)j + \alpha_h d^2_{hj} + \beta d^2_{(ik)h} + \tau \mid d^2_{(ik)j} - d^2_{hj} \mid . \tag{3}$$

Using parametric strategies common to these linkages on equations (1) and (3) we establish:

$$\tau = 0, \tag{4.a}$$

$$-\alpha_{ik}\alpha_h = -\alpha_i\alpha_k = \beta < 0. \tag{4.b}$$

For the Ward Method, condition (4.b) only holds when $N \to \infty$. Hence, if we use the test in this particular case, results will be asymptotically equivalent to those achieved with the other two methods. Then:

$$d^2_{(ikh)j} - d^2_{(ik)h} = \alpha_{ik} d^2_{(ik)j} + \alpha_h d^2_{hj} - \alpha_i d^2_{ih} - \alpha_k d^2_{kh} + \beta \mid d^2_{(ik)h} - d^2_{ik} \mid . \tag{5}$$

Setting Y_N as the distance in the aggregation level N and using equations $(4.a) - (4.b)$, we can specify equation (5) as:

$$Y_N - Y_{N-1} = \delta + \beta[Y_{N-1} - Y_{N-2}]. \tag{6}$$

Since the hierarchical distances are not variance stationary, equation (6) must be expressed in logarithms:

$$\log(Y_N) - \log(Y_{N-1}) = \delta + \beta[\log(Y_{N-1}) - \log(Y_{N-2})] + \mu, \qquad \mu \sim N(0, \sigma^2). \tag{7}$$

The usual way to choose linkage in literature has no rational justification [Blin and Cohen (1977)]. The conditions we present allow us to discriminate in some cases on the linkage to use. We have three possible situations:

a) If we get a positive β value in one of the applied method, this implies this specific linkage cannot be use properly.

b) If estimated values of β are negatives then Centroid and Ward Method can be applied, but Median Method requires β to be equal to $-1/4$.

c) If estimated values of β allow us to use more than one method, then we propose employing an information criteria to discriminate between them.

We apply the proposed test to an input–output Table of the Spanish 1985 Economy integrated by 56 sectors [see Appendix 1]. We carried out six hierarchical cluster analysis (using Euclidean distance) on the Leontief Inverse Matrix and on the technical coefficient matrix. The use of the Leontief Inverse Matrix (see Blin and Cohen, 1977) is justified for two main reasons. First of all it shows the direct and indirect inputs that each industry needs, and secondly it defines the production function at vertically integrated sector level. The applied linkages were Centroid, Median and Ward. In Appendix 2 we show the different cluster distances in each degree of aggregation. To choose the proper method, we estimate equation (7) for the six computed dissimilarity measures [see Appendix 2].[1]

Looking at the achieved results [Table 2], we can establish the only proper method to apply in that particular case is the median on the Leontief's inverse matrix. We work on the aggregation according to the linkage proposed and clustering sectors as it's shown in the dendogram in Appendix 3.

At first we work on the data using APL language. The distances and the dendogram are obtained with the SPSS/PC+ and the estimation of the autoregressive model has been achieved with the microTSP V.5.0. In Appendix 3 we can see how the dendogram performs the aggregation.

To begin with, industries named with numbers 2 and 4, and 5 and 6 are aggregated. Industries 2 and 4 represent respectively Coal and Coke; and 5 and 6 are Crude Oil and

[1] In Appendix 2, we show the distances obtained according to the method we applied. The columns noted with $\star$ mean that we are measuring them on the Leontief's inverse matrix. Names without that symbol are referred to as the coefficient matrix.

Fuel Oil. At a second stage a great number of industries are aggregated. The aggregation is carried out on well–defined sectors, *e.g.*, there are a number of industries which can be classified as Services (Public and Private), another groups can be viewed as Industries of Energy, Transports and Communications, Metals and Machinery and Non-Metals Minerals Industries, but the cluster cannot distinguish among them. At a third level the dendogram shows an aggregation of some related industries such as Paper, Paper goods, and Tobacco. At higher aggregation levels the similarity among industries became smaller, and basically it includes Manufacturing Industries.

4 Concluding Remarks

This kind of analysis let us to use an aggregation pattern to consolidate the industries of the input–output table of the Spanish economy in 1985 into sectors. This is an important tool to minimize aggregation bias, and, with the technique we have shown in this paper, we can discriminate which linkage gives the best cluster aggregation. The one with this characteristic is the Median, which gives no importance to the number of industries of each sector to be cluster. Therefore we conclude that the best aggregation method is the unweighted one.

References

Andergerg, M.R (1973): *Cluster Analysis for Applications*. Academic Press, London.

Ara, K. (1959): "The Aggregation Problem in Input–Output Analysis." *Econometrica*, 27, 257–262.

Blin, J.M. and C. Cohen (1977): "Technological Similarity and Aggregation in Input-Output Systems: A Cluster–Analytic Approach." *The Review of Economics and Statistics*, 59, 82–91.

Escudero, L.F. (1977): *Reconocimiento de Patrones*. Paraninfo.

Fisher, W.D. (1962): "Optimal Aggregation in Multi–Equation Prediction Models." *Econometrica*, 30, 744–769.

Fisher, W.D. (1969): *Clustering and Aggregation in Economics*. Johns Hopkins Press.

Hatanaka, M. (1952): "Note on Consolidation within a Leontief System." *Econometrica*, 20, 301–303.

Kossov, V. (1972): "The Theory of Aggregation in Input–Output Models," in Carter, A.P. and A. Brody (eds.): *Contributions to Input–Output Analysis*. North–Holland Publishing.

Kymn, K.O. and J.R. Norsworthy (1976): "A Review of Industry Aggregation in Input–Output Models." *The American Economist*, Spring, 5–10.

Malinvaud, E. (1954): "Aggregation Problems in Input–Output Models," in Barna, T. (ed.): *The Structural Interdependence of the Economy*. John Wiley and Sons.

Morimoto, Y. (1971): "A Note on Weighted Aggregation in Input–Output Analysis." *International Economic Review*, 12, 138–143.

Neudecker, H. (1970): "Aggregation in Input–Output Analysis: An Extension of Fisher's Method." *Econometrica*, 38, 921–926.

Sneath, P.H.A. and R.R. Sokal (1973): *Numerical Taxonomy*. W.H. Freeman Press.

Theil, H. (1957): "Linear Aggregation in Input–Output Analysis." *Econometrica*, 25, 111–122.

TABLE 1

LINKAGE	α_i	α_k	β	τ	C.1	C.2
Simple Linkage	1/2	1/2	0	-1/2	YES	YES
Complete Linkage	1/2	1/2	0	1/2	YES	YES
Centroid	$n_i/(n_i+n_k)$	$n_k/(n_i+n_k)$	$-\alpha_i\alpha_k$	0	YES	NO
Median	1/2	1/2	$-\alpha_i\alpha_k=-1/4$	0	YES	NO
Group Average	$n_i/(n_i+n_k)$	$n_k/(n_i+n_k)$	0	0	YES	YES
Weighted Aritmetic Average	1/2	1/2	0	0	YES	YES
Ward	$\frac{(n_h+n_i+n_k)n_i}{n_h(n_i+n_k)^2}$	$\frac{(n_h+n_i+n_k)n_k}{n_h(n_i+n_k)^2}$	$-\alpha_i\alpha_k$	0	YES	NO

TABLE 2

LINKAGE	MATRIX	Theoretical Values	Estimated Values	STD. ERROR
CENTROID	$(I-A)^{-1}$	$\beta < 0$	0.218	0.138
CENTROID	A	$\beta < 0$	-0.052	0.139
MEDIAN	$(I-A)^{-1}$	$\beta = -1/4$	-0.286	0.137
MEDIAN	A	$\beta = -1/4$	-0.025	0.139
WARD	$(I-A)^{-1}$	$\beta < 0$	0.677	0.010
WARD	A	$\beta < 0$	0.181	0.137

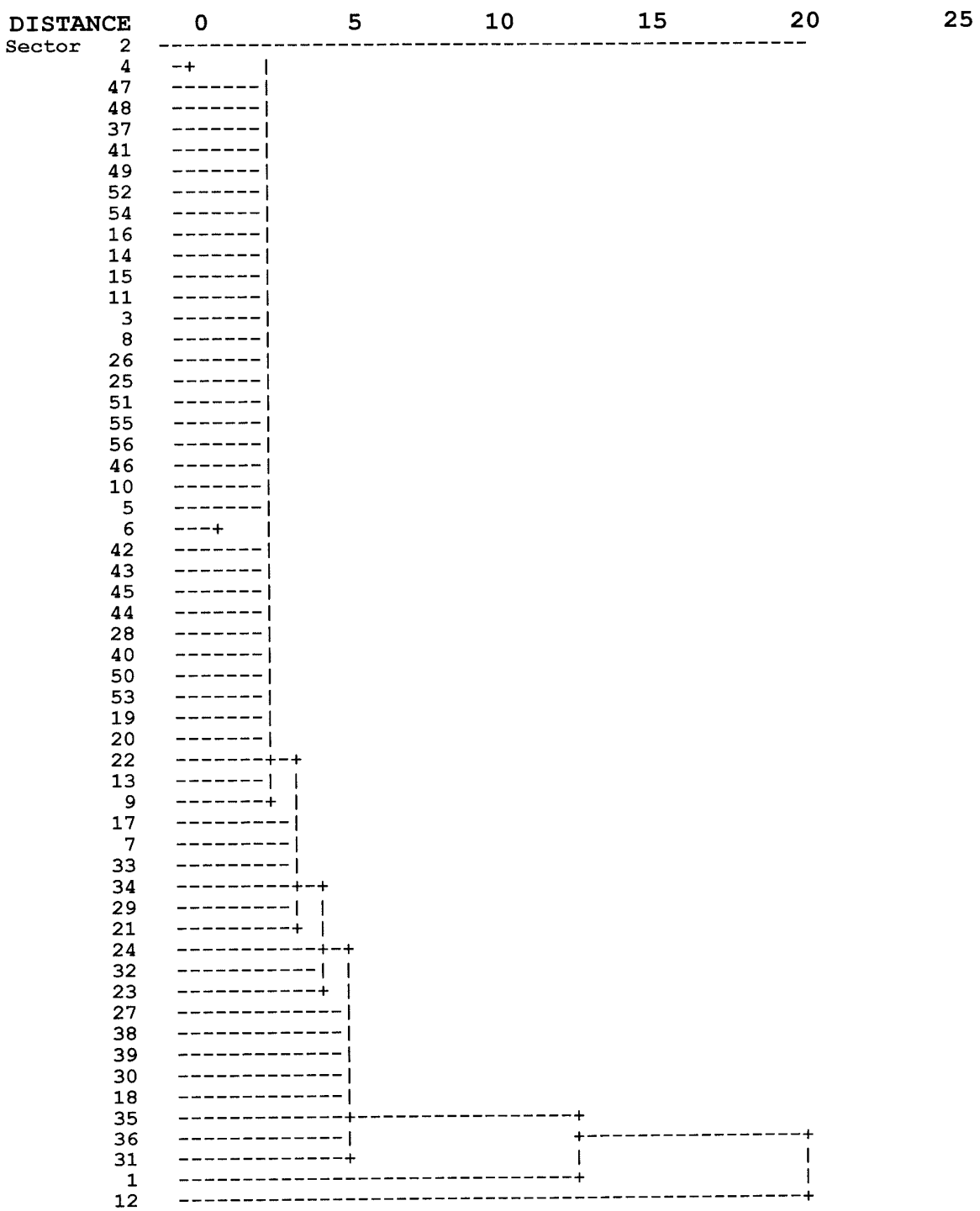
APPENDIX 1

1. Agricultural goods and Fishing.
2. Coal.
3. Lignite.
4. Coke.
5. Crude Oil.
6. Fuel Oil.
7. Natural Gas.
8. Water.
9. Electricity.
10. Manufactured Gas.
11. Nuclear Power.
12. Iron and Siderurgy.
13. Non-Iron Metals.
14. Concrete and Lime.
15. Glass.
16. Clay and Pottery.
17. Non-metalic Minerals.
18. Chemical Products.
19. Metal Products.
20. Industrial and Agricultural Machines.
21. Office Machines.
22. Electric Machines and Materials.
23. Vehicles.
24. Other Transports.
25. Meat and Canned Goods.
26. Milk.
27. Other Foods.
28. Beverages.
29. Tobacco.
30. Textiles and Clothes.
31. Leather and Shoes.
32. Wood and Wooden Furniture.
33. Paper.
34. Paper goods.
35. Rubber and plastics.
36. Other manufacturing industries.
37. Buildings.
38. Repearing and Restoring.
39. Commerce.
40. Lodging and Restaurants.
41. Railways.
42. Road Transports.
43. Ship Transports.
44. Air Transports.
45. Transport Connected Services.
46. Communications.
47. Insurances and Credits.
48. Services to Firms.
49. Hiring.
50. Private Research and Teaching.
51. Private Health.
52. Private Services.
53. Public Administration.
54. Public Research and Teaching.
55. Public Health.
56. Public Services.

APPENDIX 2: Distance Measures

obs	CENTROID*	CENTROID	MEDIAN*	MEDIAN	WARD*	WARD
1	1.086773	0.000000	1.086773	0.000000	0.543386	0.000000
2	1.237322	0.000000	1.237322	0.000000	1.162048	0.000000
3	1.509658	0.000000	1.509658	0.000000	1.916876	0.000000
4	1.606802	0.000211	1.606802	0.000211	2.720277	0.000163
5	1.498185	0.000225	1.498185	0.000231	3.552068	0.000356
6	1.407603	0.000245	1.421395	0.000291	4.408797	0.000558
7	1.286748	0.000275	1.372744	0.000323	5.324327	0.000768
8	1.230971	0.000404	1.349846	0.000404	6.261079	0.001009
9	1.215896	0.000565	1.319685	0.000617	7.230566	0.001511
10	1.182498	0.001222	1.324216	0.001326	8.202430	0.002571
11	1.164567	0.001381	1.349501	0.001490	9.176352	0.003747
12	1.141025	0.001601	1.338842	0.002106	10.15160	0.005042
13	1.129398	0.001873	1.336682	0.002225	11.14366	0.006513
14	1.126651	0.002351	1.341292	0.002351	12.14274	0.008003
15	1.118604	0.002981	1.361204	0.002981	13.14240	0.010164
16	1.120774	0.003359	1.362200	0.003260	14.14221	0.012682
17	1.073061	0.003777	1.243354	0.003239	15.14221	0.015391
18	1.083619	0.003794	1.360487	0.003777	16.14272	0.018797
19	1.102288	0.004071	1.379103	0.003784	17.14706	0.022327
20	1.102996	0.004435	1.125905	0.005511	18.15578	0.026128
21	0.888556	0.005694	1.333099	0.006369	19.16676	0.030396
22	1.111937	0.006094	1.357485	0.006495	20.18827	0.035405
23	1.032441	0.006813	1.285336	0.006813	21.21091	0.041947
24	1.070937	0.009724	1.230825	0.009578	22.23941	0.049601
25	1.090276	0.010008	1.355629	0.009736	23.27036	0.058663
26	1.101511	0.010018	1.190748	0.010018	24.30861	0.067753
27	1.095618	0.011158	1.316942	0.013857	25.35805	0.077595
28	1.144578	0.011588	1.519801	0.015105	26.42085	0.088319
29	1.023809	0.011768	1.250335	0.015313	27.49084	0.099655
30	1.050715	0.012982	1.373027	0.017654	28.58328	0.113646
31	1.061324	0.015053	1.352368	0.018181	29.67846	0.128100
32	1.182146	0.015575	1.476684	0.019087	30.81041	0.143281
33	1.211342	0.016195	1.458038	0.019683	31.95211	0.161093
34	1.222456	0.018181	1.528865	0.017346	33.12232	0.180187
35	1.244665	0.015511	1.551237	0.016911	34.30653	0.202826
36	1.245550	0.019327	1.635009	0.022928	35.51087	0.227788
37	1.302045	0.021851	1.651422	0.025011	36.73420	0.253855
38	1.318629	0.022914	1.652708	0.030001	37.97655	0.280214
39	1.344334	0.023220	1.663582	0.032214	39.27947	0.307405
40	1.357589	0.026946	1.172956	0.032009	40.58477	0.337454
41	1.367113	0.030948	1.578421	0.038599	41.89592	0.372298
42	1.385258	0.035497	1.747024	0.044018	43.21941	0.408437
43	1.034312	0.047932	1.779806	0.052133	44.59789	0.444667
44	1.419195	0.052133	1.834077	0.060594	46.02238	0.488166
45	1.422105	0.055292	1.866315	0.071353	47.46022	0.535008
46	1.469859	0.072277	1.918317	0.072277	48.93483	0.589072
47	1.549190	0.088861	1.938973	0.078300	50.42959	0.685333
48	1.569673	0.109334	0.932826	0.116193	51.93950	0.817080
49	1.619188	0.110863	1.765087	0.152690	53.47889	0.952570
50	1.045527	0.123141	1.856135	0.136648	55.10715	1.120109
51	1.019104	0.168822	1.240742	0.196916	56.82613	1.363922
52	1.020056	0.258059	1.865540	0.311220	58.84815	1.617112
53	1.459971	0.348336	1.974249	0.435513	60.87863	1.958997
54	2.833815	0.597996	3.258228	0.706700	63.91332	2.546120
55	3.538783	0.775621	4.527016	0.964452	67.38892	3.307891

APPENDIX 3: Dendrogram using Median Method. $(I-A)^{-1}$



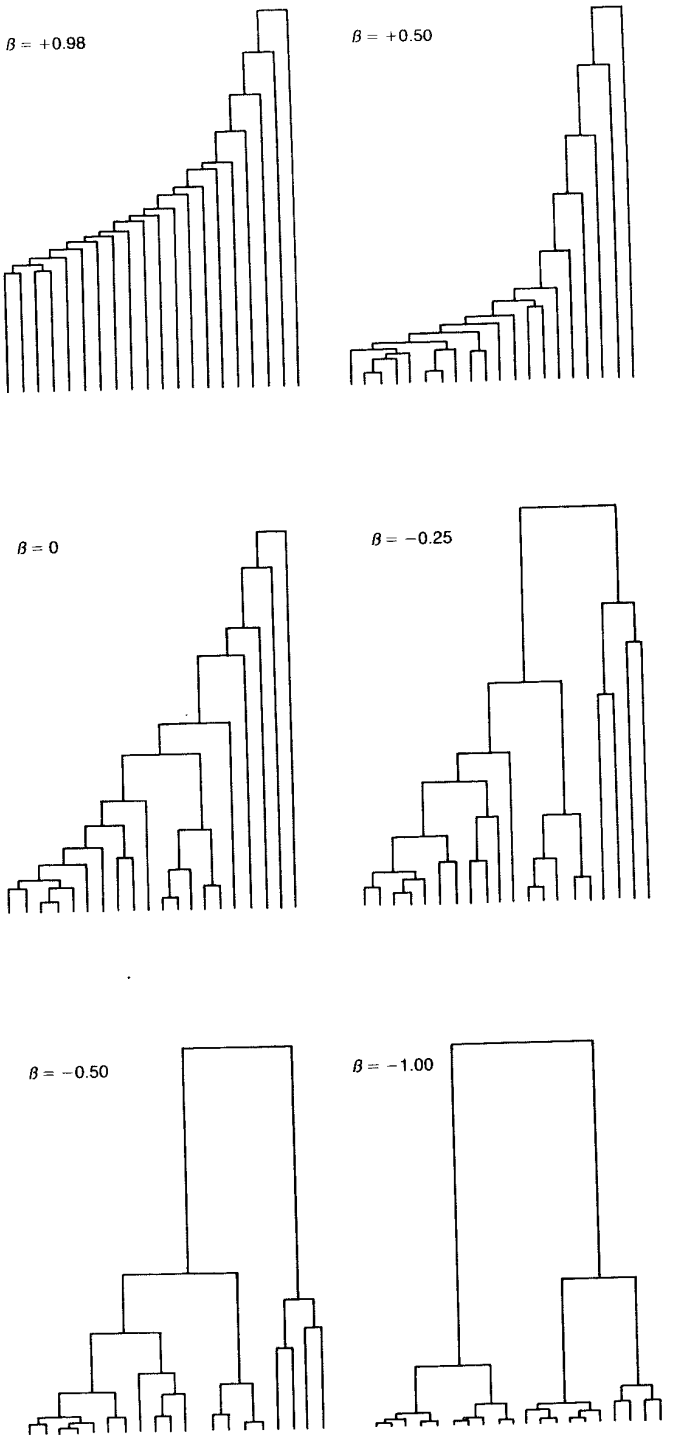


Figure 1. Dendograms.