STUDIECENTRUM VOOR ECONOMISCH EN SOCIAAL ONDERZOEK

## VAKGROEP MACRO-ECONOMIE

# Automatic ARMA Identification for a Large Number of Time Series

**E. BORGHERS**

Report 93/294

December 1993

Universitaire Faculteiten St.-Ignatius
Prinsstraat 13 - B 2000 Antwerpen
D/1993/1169/26

**Abstract**

The general known approach for the identification of ARMA models would be time consuming and not cost effective if a large number of time series is involved.

In this paper an alternative method is proposed. The method consists in the identification, within a large set of weak stationary time series, of subsets of series characterised by the same model structure but for which the model coefficients may vary. The proposed method is based on - the recursive use of - the Principal Component technique, applied to a similarity matrix.

The recursive algorithm is illustrated and commented by using simulated data. Although a lot of questions remain, these first preliminary results look very promising.

Dr. E. BORGHERS
University of Antwerp - UFSIA
Prinsstraat 13 - Room # 403
B 2000 Antwerp - Belgium

Tel.: 03/220.4131
Fax.: 03/220.4799
Email: FTE.BORGHERS.E@alpha.ufsia.ac.be

## 1. Introduction

This paper concerns the forecasting, based on ARIMA models, of a large number of time series. The main problem in forecasting a large number of time series is not the forecasting itself but the preliminary ARIMA model building process. The well-known classical approach of model identification, model estimation and model checking for each individual time series would be time-consuming and not cost-effective. It follows that this classical approach cannot be used when a large number of time series is involved.

A first way out is the use of what can be called "automatic procedures". One of the very first attempts in building a computerised automatic procedure is due to REILLY [30] and resulted in the AUTOBOX software package [36].

The first results obtained with this automatic Box-Jenkins modelling algorithm were thoroughly analysed by JENKINS [19]. ( For an overview of other comparisons see HILL & FILDES [15] and POULOS et al. [29] ). Compared with models built by a non-automatic procedure the Reilly models showed severe shortcomings. It is not surprising that the forecasts generated from these automatically built models performed rather badly.

A very recent attempt to provide an automated procedure is the module SCA-EXPERT, embedded in the SCA Statistical System [39]. This module is " ... a breakthrough in the use of expert system technology to automatically identify and estimate time series models. It employs an intelligent algorithm that is extremely reliable in identifying ARIMA models... ". However, up to now no results, practical experience with or evaluation of this relative new software have been reported in the literature.

One main disadvantage of any of these automatic modelling systems is that they can and will lead to an unpredictable number of different model structures. However well designed these systems might be, there will never be any guarantee that similar time series will be described by a common model structure. Precisely this large number of different model structures will pose serious problems in an effective monitoring of forecasts and models.

A second way out for tackling the problem of model identification, model estimation and forecasting for a large number of time series can be found in JENKINS & McLEOD [20] and BORGHERS & McLEOD [ 5]. This approach, called AUTOMOD, is based on the idea that time series describing similar situations normally need models with a similar model structure to describe their behaviour.

The AUTOMOD procedure can be presented as a three-stage procedure. First, the time series must be classified into groups that are expected to behave in similar ways. This classification can be based, for example, on product type, geographical area or sales territories, presence or absence of seasonality. Next a small representative set of time series from each group is selected for detailed analysis. Based on this detailed analysis a decision is made about the nature of a common model structure for all time series belonging to the same group. In a last stage the procedure uses a common structure to complete the model specification for each time series.

A crucial point in this AUTOMOD approach is the preliminary classification of a large number of time series into different groups. This classification is not only crucial but perhaps the weakest point in the whole approach. The objective criteria on which the classification will be based are in no way a guarantee that all time series belonging to a particular group will have an identical model structure. It is precisely this grouping of series that will be the key argument in the method presented here.

## 2. Basic Assumptions

It is assumed that $Z_t$ is a weak or covariance stationary time series, i.e. a time series for which the central tendency ( mean ) and the variability ( variance ) of the generating process is constant over time. As shown by BOX & JENKINS [ 6] this time series can be written as an ARMA(p,q) model

$$\Phi(B)\, Z_t = \Theta(B)\, a_t \qquad t = 1, 2, \ldots, T$$

where

$$\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$$

$$\Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q$$

are polynomials in the backshift operator B such that

$$B^k Z_t = Z_{t-k}$$

and $a_t$ is a Gaussian white noise process with zero mean and variance $\sigma_a^2$ .

Further, it is assumed that all the zeros of $\Phi(B)$ are outside, those of $\Theta(B)$ are on or

outside the unit circle and that $\Phi(B)$ and $\Theta(B)$ have no common factors.

It follows that the original ARMA(p,q) model can be rewritten as

$$\Pi(B)\, Z_t = a_t \qquad t = 1, 2, \ldots, T$$

where $\Pi(B)$ is an AR($\infty$) operator, defined as

$$\Pi(B) = \Phi(B)\, \Theta^{-1}\ (B)$$

or

$$\Pi(B) = 1 - \pi_1 B - \pi_2 B^2 - \ldots$$

It follows that, given initial values and known orders p and q, any ARMA model can be fully characterised by its autoregressive representation. The time series $Z_t$ is completely specified by its AR representation since any other information, except for the initial values of the series, to specify $Z_t$ is just $a_t$, which is unpredictable at time t-1.

## 3. Autoregressive Representation

Given the importance of and the crucial role played by the autoregressive representation of a weak stationary time series it is evident that attempts have been made to define the difference between time series by investigating the dissimilarity or distance between their autoregressive representation.

The ideas and implications of a distance measure for time series were first developed by CORDUAS [ 9]. The concept of a metric between ARIMA models has recently been proposed by D. PEÑA, University of Madrid, as a measure for detecting influential observations in time series. However, it was PICCOLO [28] who first proposed to use the classical Euclidean distance as a measure of dissimilarity between ARIMA models.

Allowing for a total number of N time series, i.e.

$$Z_{it} \qquad \begin{array}{l} i = 1, 2, \ldots, N \\ t = 1, 2, \ldots, T(i) \end{array}$$

where T(i) represents the total number of observations for the ith time series, the Euclidean distance between two AR representations can be defined as

$$d_{ij} = \sqrt{\sum_{l=1}^{p} (\pi_{il} - \pi_{jl})^2} \qquad i,j = 1, 2, ..., N$$

where

$$\pi_{il} \qquad i = 1, 2, ..., N$$

$$l = 1, 2, ..., p(i)$$

are the coefficients of the AR polynomial for the ith series, p(i) stands for the maximum order of the AR model for the ith series and p is defined as

$$p = \max [p(i),p(j)] \qquad i,j = 1, 2, ..., N$$

If two AR polynomials are unequal in length, the highest order is used. The missing coefficients for the shortest polynomial are set equal to zero.

It has to be stressed that the comparison of AR models, even by using the Euclidean distance, cannot be uniquely defined since any distance is essentially an arbitrary measure of diversity which satisfies well-defined and known axioms.

However, as will be seen later, the method chosen here seems to be effective with respect to the initial purpose, i.e. to discriminate between different AR structures. Furthermore, the method is statistically consistent, simple to calculate and generally applicable.

In order to facilitate the interpretation and the comparison of this Euclidean distance, it was decided to transform it. In a first stage this distance was rescaled to the zero-one range. The rescaled distance is obtained by dividing the original Euclidean distance by the maximum distance obtained, i.e.

$$d_{ij}^{*} = d_{ij}/\max( d ) \qquad i,j = 1, 2, ..., N$$

where $d_{ij}^{*}$ is the rescaled distance and max( d ) stands for the maximum distance obtained among all the $d_{ij}$'s.

Apart from being a real measure of distance the rescaled Euclidean distance can also be seen as a measure of unlikeness.

In a second stage the rescaled distance is transformed from a dissimilarity measure into a similarity measure by taking the complement value, or

$$s_{ij} = 1 - d_{ij}^{*} \qquad i,j = 1, 2, \ldots, N$$

Unlike a dissimilarity measure, which estimates the distance or the unlikeness of two objects, a similarity measure estimates the proximity or the amount of closeness between two objects.

The here defined similarity measure follows the general rule that a high value indicates much similarity and a small value indicates little similarity. In this respect $s_{ij}$ can be compared with the Pearson correlation coefficient, which is one of the most frequently used measures of similarity.

## 4. Similarity Matrix

The similarity matrix S, defined in the previous section and measuring the closeness between the AR representations of the N weak stationary time series, has some very interesting and particular characteristics. This matrix, which is an NxN matrix, has the following characteristics:

- symmetric

$$s_{ij} = s_{ji} \qquad i,j = 1, 2, \ldots, N$$

- diagonal elements equal to one

$$s_{ii} = 1 \qquad i = 1, 2, \ldots, N$$

- off-diagonal elements between zero and one, boundaries included

$$0 \leq s_{ij} \leq 1 \qquad i,j = 1, 2, \ldots, N \quad i \neq j$$

These characteristics are well known in multivariate statistics. They arise for example in the typical situation when N variables are positively correlated.

Furthermore, this matrix can be written as

$$S = LL'$$

where the matrix L is a matrix consisting of the N rescaled and orthogonal eigenvectors of S, i.e.

$$L = AD^{½}$$

or

$$S = ADA'$$

where A is the NxN matrix of eigenvectors and $D^{½}$ is the NxN diagonal matrix with the square roots of the eigenvalues of the S matrix as its diagonal elements. Without loss of generality it is assumed that the eigenvalues, and the accompanying eigenvectors, are placed in decreasing order of magnitude.

The decomposition of the similariry matrix S resembles the decomposition of the classical correlation matrix as it is used within the context of a Principal Component Analysis ( PCA ).

## 5. Principal Component Analysis

Let Y be an NxT matrix, consisting of T observations on N observed variables, each with zero mean and unit variance. Assume that each of these N variables can be written as a linear combination of N orthogonal ( uncorrelated ) unobserved hypothetical explanatory variables or components, i.e.

$$Y = LX$$

where Y is the NxT data matrix, X is the NxT matrix of explanatory components and L is an NxN matrix of coefficients, linking the correlated and observed y variables to the uncorrelated and unobserved x variables. From this it follows that the correlation matrix R, describing the intercorrelation structure among the y variables, can be written as

$$R = YY' = LXX'L' = LL'$$

By analogy with the reasoning in the previous paragraph this matrix L can also be expressed as a function of eigenvectors and eigenvalues, i.e.

$$L = AD^{½}$$

or

$$R = ADA'$$

where A is the NxN matrix of eigenvectors and $D^{½}$ is the NxN diagonal matrix with the square roots of the eigenvalues of the R matrix as its diagonal elements. Also in this situation it is assumed that, without loss of generality, the eigenvalues and accompanying eigenvectors are placed in decreasing order of magnitude.

There is a striking resemblance between the treatment of the similarity matrix S and the correlation matrix R in a PCA. Both matrices have the same characteristics. The fact that the S matrix consists of positive elements only is not exceptional since it is not so unusual to think of a correlation matrix as being a matrix with just positive correlation elements. Furthermore, both matrices can be decomposed, the decomposition being based upon the eigenvalues and the eigenvectors.

The crucial difference between the two matrices is the way the similarity and the correlation matrices are derived. The correlation matrix R, measuring the similarity between N variables, is based on and calculated from the T observations on each of these N variables. The similarity matrix S, however, measuring the similarity between the model structure of N time series, is not based on the N time series but on a transformation of the dissimilarity between the estimated AR representations of each of these time series.

It has to be noticed that the principal component analysis is a mathematical technique which does not require to specify any underlying statistical model to explain the " error " structure. In particular, no specific assumption has to be made about the probability distribution of the original data.

## 6. Factor-Pattern Matrix

Principal component analysis consists in finding an orthogonal set of unobserved constructs, called components, to describe the original and observed correlated data. These components are derived in decreasing order of importance. The criterion used in measuring this importance is based upon the relative contribution of each of these components in the reconstruction of the similarity and the correlation matrix. It is hoped that the first few components will account for most of the variance in the original data so that the effective dimensionality of the data can be reduced.

However, principal component analysis is not just a technique to reduce the dimension of the observed variables. In many situations one tries to interpret these unobserved components in a meaningful way. This interpretation is based upon the columns of matrix L, consisting of the rescaled eigenvectors. In many respects the coefficients of this matrix, called factor-pattern matrix or matrix of factor loadings, are like regression coefficients. They predict the output variables y from some input or explanatory x variables. It follows that this matrix can be seen as consisting of weights to be assigned to the principal components in deriving the observed variables as linear combinations of these principal components.

One common procedure to interpret these components is to split the variables into groups which are associated with particular components. These groups often have the property that variables within the same group are highly correlated while variables in different groups have low correlations. Frequently these groups can even be found by looking directly at the correlation matrix and grouping the variables in a way based on common-sense.

Using a principal component analysis to analyse the similarity matrix S looks promising given some well-known characteristics of this similarity matrix. These specific characteristics, i.e.

- symmetry
- diagonal elements equal to one
- off-diagonal elements between zero and one, boundaries included

are well known in multivariate statistics ( See for example CHATFIELD & COLLINS [ 8, p. 72] ). It is a common situation that arises when all variables are positively correlated.

If these characteristics are present, the loadings for the first principal component, this is the component based on the eigenvector corresponding to the largest eigenvalue, will consist of only positive elements. Since a principal component analysis must be situated in a deterministic context it is hard to see how to use any criterion to discriminate among all these positive loadings. From this it follows that there is little point in looking at the loadings for this first component.

Although it will not be used any further directly, indirectly this first component will play an interesting and crucial role. This component can be seen as some sort of weighted average of AR representations. It can be regarded as a general mean measure of size.

Due to the orthogonality properties of the eigenvectors, and thus of the principal components, at least one of the elements of the second eigenvector must be negative and at least one must be positive. For this reason it might be expected that this second principal component will be of more interest than the first in discriminating among the factor loadings.

This second component, i.e. the component corresponding to the second largest eigenvalue, can be seen as a first correction for the general mean level obtained as the contribution of the first component. The main purpose of this paper is to investigate how this second component performs in correcting this general mean level, taking into account the specific pattern of the similarity matrix, induced by the presence of possible homogeneous groups of different model structures.

## 7. Technical and Practical Problems

Before conducting some experiments a few problems have still to be solved.

### Order Criteria for AR Schemes

A first problem that has to be solved is the determination of the order of each of the autoregressive models.

A well-known criterion to determine the optimal order for an AR process is given by AKAIKE [ 1]. Although this criterion is based on general principles of information theory, its interpretation is rather obvious. In this criterion, given by

$$AIC(k) = \log \sigma^2(k) + \frac{2k}{T}$$

the first term is a measure of how well the AR model fits the data, while the second term is a penalty for the number of parameters in the model. As the optimal order for the AR model that number for k is chosen for which AIC reaches a global minimum.

Another criterion that could have been used for determining the order of an AR process is the CAT criterion, due to PARZEN [27]. Instead of assuming that the time series is an AR(p) process, PARZEN considered the problem of determining the order p of an AR process that approximates in some optimal way the behaviour of an arbitrary covariance or weak stationary time series.

In practice, identical conclusions are usually obtained from AIC and CAT. Besides, it has been shown by BHANSALI [ 4] that the AIC and the CAT criterion are asymptotically equivalent.

It was decided to use the AIC criterion to determine the order of eachof the AR models. The practical use consisted in the fitting of autoregressive models of successively higher orders. For each order the output consisted of the estimate of the Π-coefficients, the innovation or residual variance and the corresponding AIC statistic.

Estimation Method

A second problem is the decision about the choice of the estimation procedure to be used in estimating each of the AR models.

As shown by TIAO & TSAY [33] , for ARMA models the ordinary least squares estimates of the autoregressive parameters in autoregressive fittings are not always consistent. In TSAY & TIAO [34] an alternative and iterative procedure is proposed that yields consistent least squares estimates for AR parameters.

Despite these arguments it was decided to use the ordinary least squares estimation method. The main reason for still using this estimation method is the fact that any alternative and appropriate estimation method yielding "better" characteristics for the estimates, in some way is based on assumptions about the possible model structure.

Precisely these a priori assumptions are conflicting with the general purpose of this paper. Since it is one of the main intentions of this paper to investigate the possibility of identifying subsets of time series sharing the same model structure, no preliminary and/or a priori assumptions about this model structure can be made.

Eigenvalues

A third and last problem is to determine the number of eigenvalues that have to be extracted.

When plotting the eigenvalues, i.e. plotting the total variance associated with each component, this plot almost always shows a distinct break between the steep slope of the large components and the gradual trailing-off of the rest of the components. Following CATTELL [ 7] this gradual trailing-off is called the scree because it resembles the rubble that forms at the foot of a mountain. Experimental evidence indicates that the scree begins at the kth component, where k is the true number of components.

Another criterion that can be used is to retain only components with eigenvalues greater than one. In essence this criterion, which is due to KAISER [21], is like saying that, unless a component extracts at least as much as the equivalent of one original variable, it is dropped. Components with a variance ( eigenvalue ) less than one are no better than a single variable, since each variable has a variance of one.

Based on simulation studies, such as HARKSTIAN et al. [12] , the CATTELL and the KAISER criteria have been compared. From these studies it follows that the KAISER criterion sometimes retains too many components while the CATTELL technique sometimes retains too few. However, both criteria do quite well under normal conditions, that is, when there are relatively few components and many variables. For further references on these and other criteria see JACKSON [17, pp. 41-52 ] and ZWICK & VELICER [35].

Cluster Analysis

In presenting the results of the experiments use will be made of Cluster Analysis. The basic aim of Cluster Analysis is to find the natural groupings, if any, of a set of objects or points. More formally, Cluster Analysis aims to allocate a set of individuals to a set of mutually exclusive, exhaustive groups such that individuals within a group are similar to one another while individuals in different groups are dissimilar. This set of groups is usually called a partition.

The groups forming a partition may be subdivided into smaller sets or grouped into larger sets, so that one eventually ends up with the complete hierarchical structure of the given set of individuals. This structure is often called a hierarchical tree. One way of visually representing the steps in a hierarchical clustering solution is by a two-dimensional display or diagram called dendrogram. This dendrogram identifies the clusters being combined.

One of the decisions that has to be made before using the tree clustering is how to determine the distance between clusters. To put it differently, a rule needs to be established for deciding when two clusters are to be amalgamated or linked. Many methods can be used to decide which objects or clusters should be combined at each step. For a review and a comparison of these methods see JARDINE & SIBSON [18], EVERITT [11, Chapter 4 ] and EVERITT [10].

The methods that will be used in presenting the results of the experiments are :

- Unweighted Pair-Group Centroid

  In the "Unweighted Pair-Group Method using the Centroid Average" the distance between two clusters is determined as the difference between centroids. This centroid can be seen as the centre of gravity for the respective clusters.

  If there are, or one suspects there to be, considerable differences in cluster sizes, this method has to be replaced by the "Weighted" version of this method. This weighted method is identical to the previous one except that weighting is introduced into the computations to take into consideration differences in cluster sizes.

- Unweighted Pair-Group Average

  In the "Unweighted Pair-Group Method using Arithmetic Averages" or the "Average Linkage Between Groups Method" the distance between two clusters is defined as the average of the distances between all pairs of points in which one point of the pair is from each of the clusters.

  This method differs from other linkage methods in that it uses information about all pairs of distances, not just the nearest ( Single Linkage Method ) or the furthest ( Complete Linkage Method ).
  Also for this method the weighted version must be used when the cluster sizes are suspected to be greatly uneven. This method is very efficient when the objects form natural distinct "clumps". However, it performs equally well with elongated "chain" type clusters.

## 8. Experiments

The procedure used in each of the reported experiments can be summarised in the following algorithmic formulation:

- For each given model structure, with given coefficients and residual variance, a number of weak stationary time series of given length was simulated.

- For each of these time series an AR-model was estimated by the method of ordinary least squares. The order for each of these AR-models was determined by the AIC criterion. The maximum allowable lag was set equal to 20 lags.

The practical use consisted in the fitting of autoregressive models of successively higher orders by the method of least squares and realised through the Householder transformation.

- After each set of Π-polynomials, belonging to and estimated from the simulated time series for a particular given model structure, the theoretical Π-polynomial for this particular model was inserted.

- Based on the estimated and theoretical Π-polynomials a distance matrix D was constructed. As a measure of distance the Euclidean distance was used.

The main advantage of inserting the theoretical Π-polynomials is that the Π-coefficients, estimated from the simulated series, can easily be compared with the a priori given coefficients.

The effect of this insertion is that the dimension of the distance matrix D is augmented by the number of a priori given model structures. Results and output for these theoretical cases will be marked by a right arrow ( -> ).

- By first scaling the individual elements of this distance matrix D between zero and one and taking the complement, this distance matrix D was then transformed from a dissimilarity matrix into a similarity matrix S. It is this similarity matrix that will be used for a Principal Component analysis.

All the experiments are based on simulated data. In all of the experiments the length mentioned for each of the time series is the length after the first 50 simulated observations were omitted from the created sequence. The reason for this is that when simulating time series, the first simulated data are often used in the calculation of the subsequent simulated values. In such cases, the recursive relationship being used may be more valid later in the simulated sequence. Therefore, more data were created than the number actually desired. When all the data points were simulated, the excess of observations were removed from the beginning of the sequence.

8.1. Experiment 1

In this first experiment time series were simulated for the following model structures:

| Model | Structure | Coefficient |
|---|---|---|
| 1 | AR(1) | 0.9 |
| 2 | AR(1) | 0.7 |
| 3 | AR(1) | -0.9 |
| 4 | AR(1) | -0.7 |
| 5 | MA(1) | 0.9 |
| 6 | MA(1) | 0.7 |
| 7 | MA(1) | -0.9 |
| 8 | MA(1) | -0.7 |

For each of these model structures five series were simulated, each series consisting of 200 observations. The variance for the normal distributed residuals was set equal to one.

Taking into account the insertion of the theoretical autoregressive polynomials, it follows that the resulting distance and similarity matrices are of the order 48 by 48, i.e. eight model structures with five estimated and one theoretical Π-polynomial for each model. The thus derived similarity matrix can be partitioned as an 8 by 8 block matrix, the diagonal blocks giving the similarities within the same model structure and the off-diagonal blocks describing the similarities between simulated time series, and theoretical models, belonging to different model structures.

In order to conserve space, not the whole similarity matrix is given but the mean similarity matrix. This matrix, represented in Table 8.1 , is derived from the large similarity matrix by calculating the ( arithmetic ) mean similarity for each block of the partitioned similarity matrix. In calculating the mean value for the diagonal blocks, the diagonal elements, which are equal to one, are left out from the calculations.

TABLE 8.1 : MEAN SIMILARITY MATRIX (°)

| Model | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | 89 | | | | | | | |
| 2 | 90 | 98 | | | | | | |
| 3 | 41 | 49 | 97 | | | | | |
| 4 | 48 | 56 | 91 | 94 | | | | |
| 5 | 21 | 27 | 49 | 49 | 73 | | | |
| 6 | 45 | 52 | 78 | 80 | 63 | 91 | | |
| 7 | 57 | 56 | 29 | 35 | 15 | 35 | 75 | |
| 8 | 72 | 73 | 42 | 48 | 26 | 49 | 74 | 86 |

(°) Mean similarities multiplied by 100 and rounded.

From Table 8.1 a few interesting conclusions can be drawn. As could be expected, the similarities within each model structure are substantially higher than the similarities between different model structures. As a second point of interest it must be mentioned that the mean similarity between AR-models with opposite sign for their coefficient, i.e. model structures 1 and 2 against model structures 3 and 4, is higher than the mean similarity between MA-models with opposite sign for their coefficient, i.e. model structures 5 and 6 against model structures 7 and 8. This is very probably due to the fact that the MA-models are approximated by truncated AR-models. A third and last remark about this mean similarity matrix can be formulated about the comparison of AR- and MA-models. Since this similarity matrix is based on AR and inverted MA specifications it is not surprising that the AR-models are showing a higher similarity with MA-models with an opposite sign for their coefficient than MA-models with the same sign for their coefficient, i.e. model structures 1 and 2 against model structures 7 and 8 or model structures 3 and 4 against model structures 5 and 6.

The first result from using the whole similarity matrix for the Principal Component analysis consists in the eigenvalues. From these eigenvalues, for which the first ten out of the total of 48 are given in TABLE 8.2 , and using the Sree-test as a criterion, it can be seen that retaining four eigenvalues will be a good compromise.

TABLE 8.2 : EIGENVALUES AND ( CUMULATIVE ) PERCENTAGES

| | Eigenvalue | Percentage | Cumul.Perc. |
|---|---|---|---|
| 1 | 27.8447 | 58.0098 | 58.0098 |
| 2 | 8.6569 | 18.0353 | 76.0451 |
| 3 | 2.9264 | 6.0966 | 82.1418 |
| 4 | 2.5856 | 5.3867 | 87.5285 |
| 5 | .9160 | 1.9082 | 89.4367 |
| 6 | .8437 | 1.7577 | 91.1944 |
| 7 | .3719 | .7748 | 91.9692 |
| 8 | .3588 | .7476 | 92.7168 |
| 9 | .3038 | .6329 | 93.3497 |
| 10 | .2844 | .5925 | 93.9422 |

The factor loadings, corresponding to the three largest of these eigenvalues, are given in TABLE 8.3. As could be expected, the loadings for the first principal component are all positive. A second and even more important remark about these principal components is the pattern of the factor loadings obtained for the second and third component. The least thing one can say about this pattern is that it is very systematic, regular and consistent.

A careful analysis of this pattern reveals that, by just using the sign of the factor loadings, the whole set of time series can be split up into subsets. Based on the first and second principal component, and using only the sign of the factor loadings, the following partition of the 48 cases can be derived:

GROUP 1 : Model 3, 4, 5, 6
GROUP 2 : Model 1, 2, 7, 8

TABLE 8.3 : PRINCIPAL COMPONENTS (°)

| Model | Series | 1 | 2 | 3 | Model | Series | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 81 | -43 | -19 | 5 | 25 | 70 | 47 | 22 |
| 1 | 2 | 77 | -42 | -16 | 5 | 26 | 47 | 46 | 52 |
| 1 | 3 | 78 | -37 | -23 | 5 | 27 | 60 | 46 | 39 |
| 1 | 4 | 76 | -51 | - 3 | 5 | 28 | 49 | 44 | 50 |
| 1 | 5 | 77 | -47 | -13 | 5 | 29 | 45 | 47 | 51 |
| 1 | -> 6 | 80 | -45 | -19 | 5 | -> 30 | 43 | 42 | 47 |
| 2 | 7 | 84 | -38 | -19 | 6 | 31 | 85 | 38 | 2 |
| 2 | 8 | 85 | -35 | -19 | 6 | 32 | 80 | 39 | 16 |
| 2 | 9 | 84 | -39 | -20 | 6 | 33 | 83 | 40 | 5 |
| 2 | 10 | 85 | -37 | -19 | 6 | 34 | 80 | 42 | 14 |
| 2 | 11 | 83 | -41 | -20 | 6 | 35 | 83 | 38 | 12 |
| 2 | -> 12 | 84 | -38 | -20 | 6 | -> 36 | 80 | 43 | 17 |
| 3 | 13 | 82 | 41 | -24 | 7 | 37 | 54 | -43 | 29 |
| 3 | 14 | 80 | 48 | -21 | 7 | 38 | 71 | -46 | 27 |
| 3 | 15 | 79 | 48 | -21 | 7 | 39 | 65 | -48 | 29 |
| 3 | 16 | 79 | 48 | -21 | 7 | 40 | 59 | -44 | 31 |
| 3 | 17 | 80 | 48 | -21 | 7 | 41 | 80 | -43 | 22 |
| 3 | -> 18 | 80 | 48 | -21 | 7 | -> 42 | 44 | -43 | 30 |
| 4 | 19 | 82 | 36 | -14 | 8 | 43 | 69 | -43 | 31 |
| 4 | 20 | 85 | 37 | -20 | 8 | 44 | 79 | -43 | 21 |
| 4 | 21 | 85 | 40 | -21 | 8 | 45 | 78 | -34 | 14 |
| 4 | 22 | 86 | 36 | -20 | 8 | 46 | 80 | -44 | 13 |
| 4 | 23 | 84 | 42 | -21 | 8 | 47 | 81 | -42 | 19 |
| 4 | -> 24 | 84 | 41 | -21 | 8 | -> 48 | 82 | -43 | 16 |

(°) Elements multiplied by 100 and rounded.

Since also the third eigenvalue can be considered as being "significant" the partitioning of the 48 cases, using the sign of the factor loadings, can also be based on the second and third principal component. Partitioning all the time series in this way leads to an even more detailed result, i.e.

GROUP 1 : Model 7, 8
GROUP 2 : Model 5, 6
GROUP 3 : Model 1, 2
GROUP 4 : Model 3, 4

An explanation for these rather surprising and unexpected results can be found in the meaning and functioning of the Principal Component technique. As it was shown in section 5, each of the components is contributing in explaining or reconstructing the similarity matrix. The relative contribution of each component is directly related to its corresponding eigenvalue.

Since the weights for the first component, i.e. the factor loadings or the elements of the rescaled eigenvector, are all positive, the contribution of this first component can be seen as a general mean level for each of the elements of the similarity matrix. It is a first step in the complete reconstruction of the similarity coefficients.

The explanatory power of the remaining components, i.e. from the second to the last component, will consist in gradually correcting this general mean value. This correcting process will stop with the contribution of the last component and will result in the fully reconstructed similarity matrix.

However, the smaller the eigenvalues, the smaller the relative explanatory power and the more likely this explanatory power will be used to correct individual and isolated similarity coefficients.

It follows that the most important drawback of this approach is a conflict situation. On the one hand there is the tendency to derive a highly detailed partitioning of the set of time series. On the other hand there is the difficulty that this can only be achieved by using additional components, for which the eigenvalues have to pass the controversial Sree-test and for which the problem of interpretation will become sharper the smaller the explanatory power, and the eigenvalues of these additional components, will be.

In order to avoid this disadvantage and the conflict situation, another approach will be proposed and illustrated in a second experiment.

Before proceeding, however, with this second experiment, the results obtained will be compared with those obtained with an alternative technique. In order to get a better and alternative insight into the working of this partitioning mechanism, a Cluster Analysis was carried out.

The first method used in clustering the 48 cases was the Centroid method. The graphical results are presented in Figure 8.1. The main conclusion that can be drawn from these results is that, without taking the value of the model coefficients into account, the

method seems to succeed in classifying each model structure in a separate cluster. Each of the four identified clusters consists in an homogeneous group of time series, described by the same model structure, although the coefficients of these models may vary.

As a second clustering solution the Average Linkage method was used. The graphical results, represented as a dendrogram, are given in Figure 8.2. Compared with the results obtained for the previous method one can conclude that also this method succeeds in clustering the AR-models in distinct clusters but that the clusters obtained for the MA-models are not completely homogeneous anymore. As a concluding remark one could say that this Average Linkage method seems to be much more sensitive to the coefficients of the models, certainly for the MA-models, than the Centroid method.

FIGURE 8.1 : DENDROGRAM CENTROID METHOD

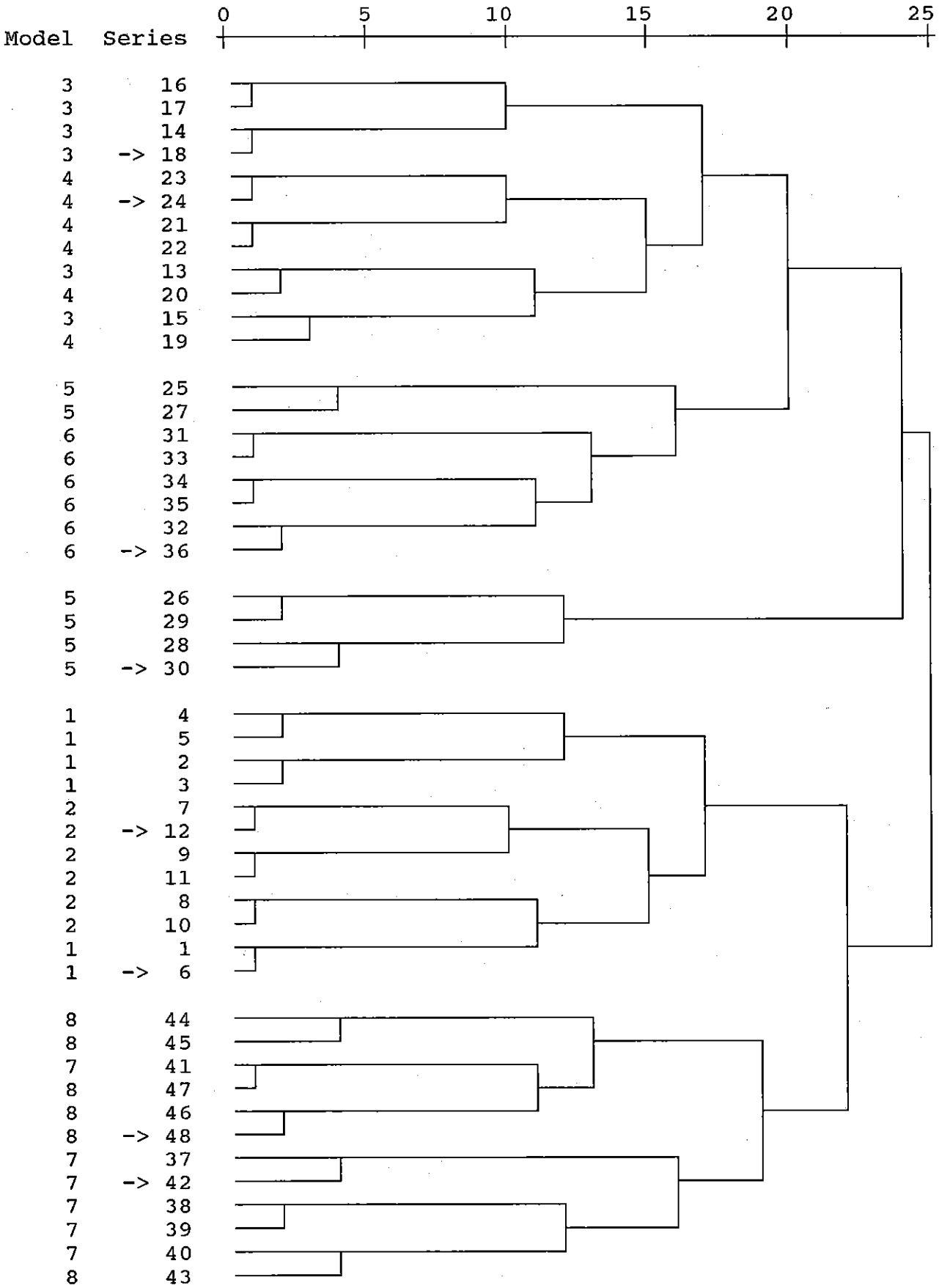
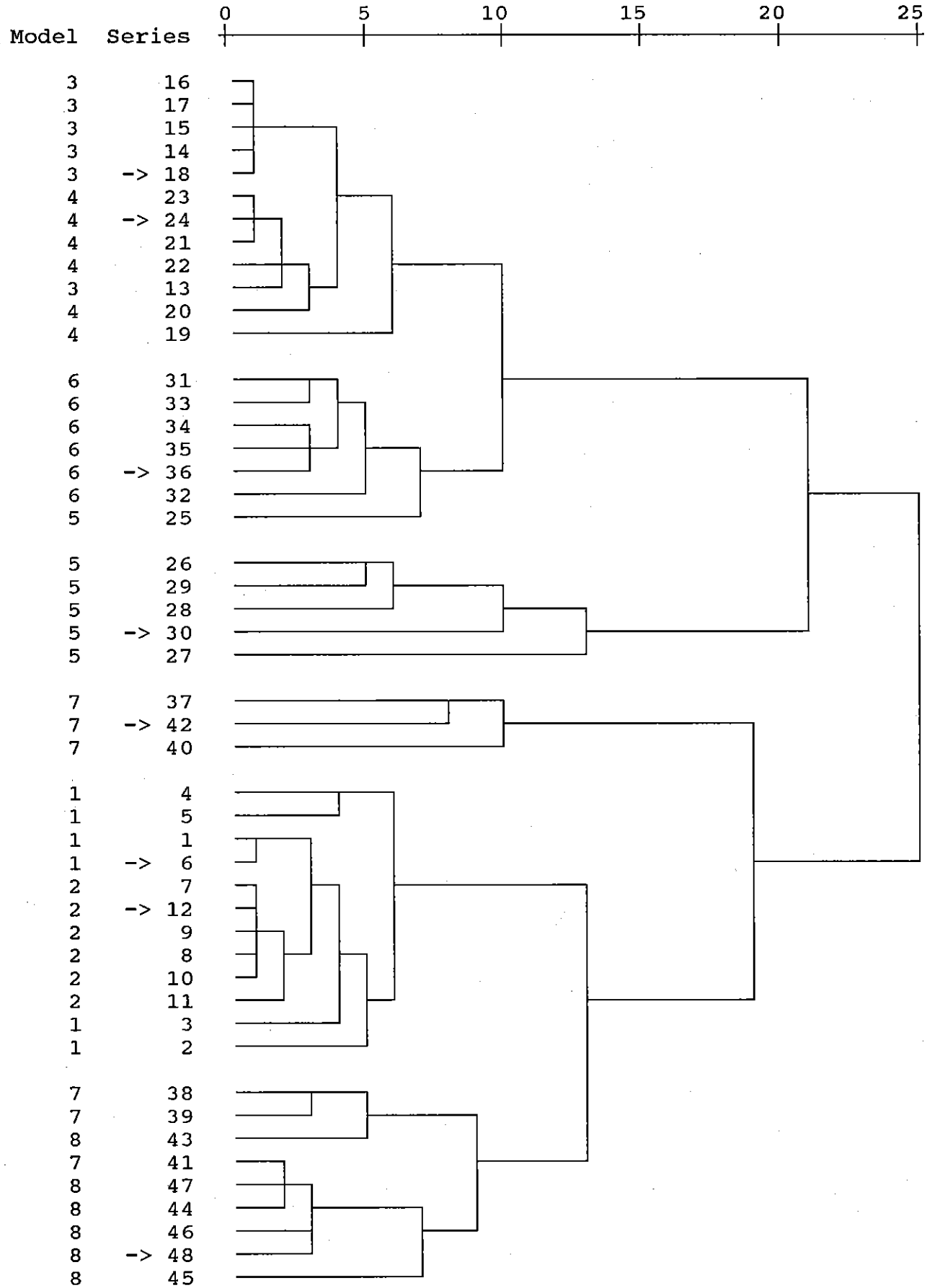


FIGURE 8.2 : DENDROGRAM AVERAGE LINKAGE METHOD



## 8.2. Experiment 2

For this second experiment time series were simulated for the following model structures:

| Model | Structure | | Coefficients | |
|---|---|---|---|---|
| 1 | (1,0) | (0,0) | 0.5 | |
| 2 | (2,0) | (0,0) | 0.5 | -0.5 |
| 3 | (0,0) | (1,0) | 0.5 | |
| 4 | (1,0) | (1,0) | 0.5 | 0.5 |
| 5 | (0,1) | (0,0) | 0.5 | |
| 6 | (0,2) | (0,0) | 0.4 | 0.4 |
| 7 | (0,0) | (0,1) | 0.5 | |
| 8 | (0,1) | (0,1) | 0.5 | 0.5 |
| 9 | (1,0) | (0,1) | 0.5 | 0.5 |
| 10 | (0,1) | (1,0) | 0.5 | 0.5 |

where the usual (p,q)(P,Q) notation is used to characterise the models.

Six out of a total of ten model structures are seasonal models ( models 3 and 7 ) or mixed nonseasonal-seasonal models ( models 4, 8, 9 and 10 ). The seasonal time span for the seasonal component is four observations. Also for this experiment five series, each consisting of 200 observations, were simulated for each of the model structures. The variance of the normal distributed residuals was set equal to one.

Taking the insertion of the theoretical autoregressive polynomials into account, it follows that the distance and the similarity matrices are of the order 60 by 60 while the mean similarity matrix will be a 10 by 10 matrix. Since the ten models can be partitioned into three groups, i.e. AR-models ( models 1, 2, 3 and 4 ), MA-models ( models 5, 6, 7 and 8 ) and ARMA-models ( model 9 and 10 ) this mean similarity matrix, given in Table 8.4, can be represented as a 3 by 3 block matrix.

TABLE 8.4 : MEAN SIMILARITY MATRIX (°)

| Model | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 94 | | | | | | | | | |
| 2 | 71 | 88 | | | | | | | | |
| 3 | 62 | 54 | 88 | | | | | | | |
| 4 | 69 | 59 | 70 | 82 | | | | | | |
| 5 | 42 | 42 | 54 | 32 | 89 | | | | | |
| 6 | 34 | 41 | 35 | 23 | 62 | 75 | | | | |
| 7 | 60 | 52 | 45 | 38 | 57 | 50 | 80 | | | |
| 8 | 33 | 33 | 31 | 17 | 63 | 65 | 60 | 74 | | |
| 9 | 65 | 57 | 36 | 39 | 34 | 32 | 64 | 36 | 80 | |
| 10 | 36 | 38 | 58 | 34 | 71 | 46 | 40 | 41 | 22 | 82 |

(°) Mean similarities multiplied by 100 and rounded.

This mean similarity matrix shows some interesting features. One of the most striking of these features is the behaviour of the model structures 4 and 8, i.e. models consisting of a multiplicative nonseasonal and seasonal model component. If both model components of these mixed models are of the autoregressive form, the similarities with the other autoregressive models are much more pronounced than with the moving average models. If both model components are of the moving average type, the similarities with the other moving average models are larger than the similarities with the autoregressive models.

TABLE 8.5 : EIGENVALUES AND ( CUMULATIVE ) PERCENTAGES

| | Eigenvalue | Percentage | Cumul.Perc. |
|---|---|---|---|
| 1 | 30.7670 | 51.2783 | 51.2783 |
| 2 | 8.3223 | 13.8705 | 65.1489 |
| 3 | 5.5530 | 9.2551 | 74.4039 |
| 4 | 2.3782 | 3.9636 | 78.3675 |
| 5 | 1.9731 | 3.2885 | 81.6560 |
| 6 | 1.0819 | 1.8032 | 83.4592 |
| 7 | .9835 | 1.6391 | 85.0983 |
| 8 | .7471 | 1.2452 | 86.3435 |
| 9 | .6518 | 1.0864 | 87.4299 |
| 10 | .5382 | .8970 | 88.3269 |

In Table 8.5 the ten largest eigenvalues of the whole similarity matrix are tabulated. The large number of eigenvalues passing the Scee-test is not surprising and could be expected since for this second experiment the total number of distinct model structures is equal to 10.

Proceeding in the same line as in the previous experiment the factor loadings, corresponding to the three largest of the eigenvalues, were calculated. The results are given in Table 8.6. Based on the first and second principal component, and using only the sign of the factor loadings, the following classification of the 10 model structures can be derived:

GROUP 1 : Model 1, 2, 3, 4, 9
GROUP 2 : Model 5, 6, 7, 8, 10

If also the third component is used the partitioning of all the time series will lead to the following even more detailed result:

GROUP 1 : Model 3, 4
GROUP 2 : Model 1, 2, 9
GROUP 3 : Model 5, 10
GROUP 4 : Model 6, 7, 8

In both of these classification schemes only one single time series, i.e. series 38, was erroneously classified. Specific analysis of this particular simulated series revealed that rather than speaking about erroneous classification this exception could be explained by a poor simulation of the series.

TABLE 8.6 : PRINCIPAL COMPONENTS (°)

| Model | Series | 1 | 2 | 3 | Model | Series | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 80 | -45 | 13 | 6 | 31 | 58 | 43 | 15 |
| 1 | 2 | 83 | -39 | 10 | 6 | 32 | 72 | 47 | 2 |
| 1 | 3 | 77 | -48 | 14 | 6 | 33 | 58 | 42 | 21 |
| 1 | 4 | 80 | -38 | 9 | 6 | 34 | 75 | 44 | 0 |
| 1 | 5 | 81 | -43 | 12 | 6 | 35 | 68 | 47 | 11 |
| 1 | -> 6 | 81 | -44 | 12 | 6 | -> 36 | 56 | 50 | 14 |
| 2 | 7 | 75 | -34 | 11 | 7 | 37 | 76 | 15 | 36 |
| 2 | 8 | 84 | - 7 | 0 | 7 | 38 | 73 | - 3 | 28 |
| 2 | 9 | 78 | -38 | 12 | 7 | 39 | 75 | 8 | 31 |
| 2 | 10 | 72 | -39 | 13 | 7 | 40 | 79 | 17 | 38 |
| 2 | 11 | 72 | -35 | 12 | 7 | 41 | 84 | 7 | 25 |
| 2 | -> 12 | 76 | -33 | 11 | 7 | -> 42 | 77 | 13 | 39 |
| 3 | 13 | 77 | -14 | -45 | 8 | 43 | 75 | 42 | 6 |
| 3 | 14 | 71 | -30 | -45 | 8 | 44 | 66 | 51 | 27 |
| 3 | 15 | 78 | -18 | -42 | 8 | 45 | 48 | 58 | 25 |
| 3 | 16 | 73 | -23 | -43 | 8 | 46 | 64 | 50 | 24 |
| 3 | 17 | 79 | -19 | -38 | 8 | 47 | 65 | 38 | 27 |
| 3 | -> 18 | 76 | -18 | -45 | 8 | -> 48 | 61 | 53 | 25 |
| 4 | 19 | 67 | -49 | -24 | 9 | 49 | 71 | -34 | 37 |
| 4 | 20 | 59 | -48 | -28 | 9 | 50 | 66 | -29 | 45 |
| 4 | 21 | 69 | -43 | -31 | 9 | 51 | 65 | -21 | 49 |
| 4 | 22 | 66 | -50 | -21 | 9 | 52 | 58 | -18 | 49 |
| 4 | 23 | 69 | -51 | -19 | 9 | 53 | 70 | -15 | 51 |
| 4 | -> 24 | 67 | -50 | -28 | 9 | -> 54 | 65 | -23 | 50 |
| 5 | 25 | 81 | 33 | -20 | 10 | 55 | 59 | 27 | -47 |
| 5 | 26 | 72 | 47 | -25 | 10 | 56 | 58 | 47 | -34 |
| 5 | 27 | 75 | 50 | -14 | 10 | 57 | 70 | 29 | -46 |
| 5 | 28 | 73 | 46 | -25 | 10 | 58 | 70 | 21 | -50 |
| 5 | 29 | 80 | 43 | -10 | 10 | 59 | 73 | 32 | -42 |
| 5 | -> 30 | 79 | 45 | -18 | 10 | -> 60 | 65 | 24 | -51 |

(°) Elements multiplied by 100 and rounded.

Since the maximum number of groups that can be identified, by only using the sign of the factor loadings for the first three principal components, is restricted to four, the alternative approach will be used. One of the basic characteristics of this proposed alternative method is that a set of time series is split up into two subsets by only using

the sign of the elements of the second component. The key point of this procedure, however, is that this splitting-up technique is not only used once but repeatedly on an ever smaller subset.

Formulated as a recursive algorithm, this alternative heuristic approach can be summarised as follows:

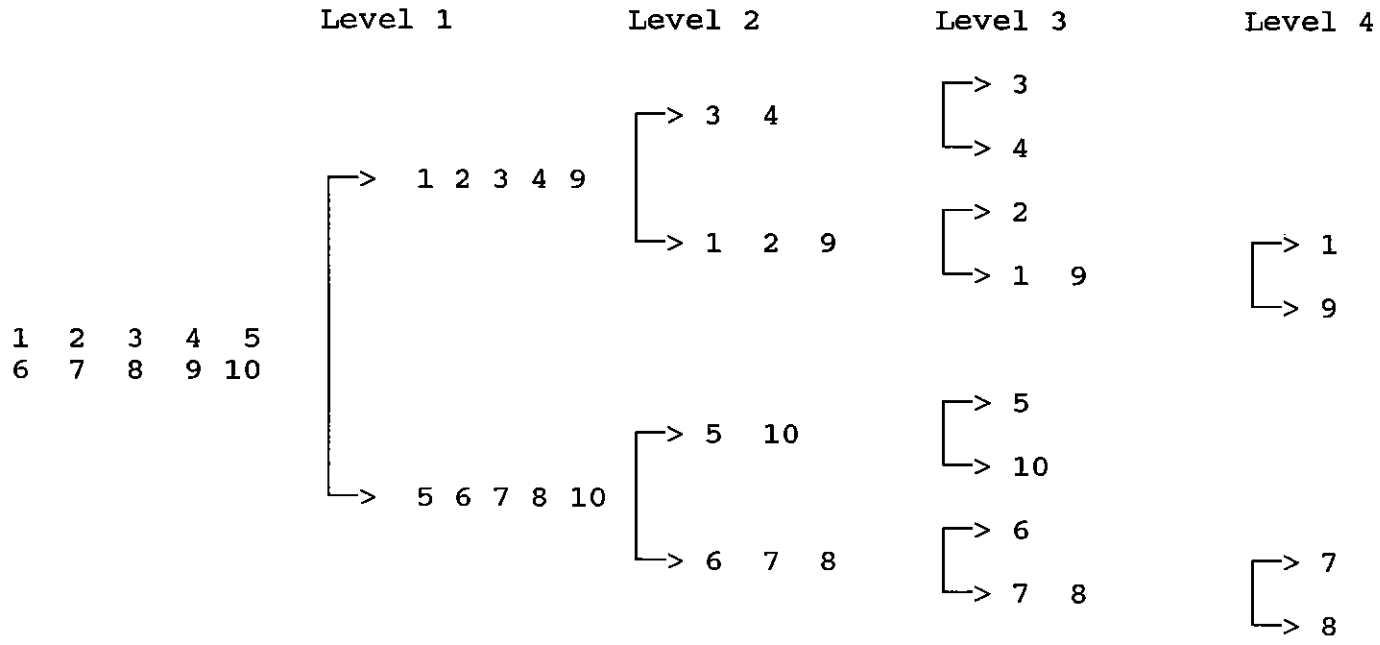
```
MODULE : Splitting Up ( Set of Time Series)
IF Stopping Rule
  THEN Stop
  ELSE Calculate Two Component Model
       Splitting Up ( Time Series with Pos. Weights for 2nd Component )
       Splitting Up ( Time Series with Neg. Weights for 2nd Component )
```

A workable definition for the stopping rule used in this algorithm could be formulated as: Only the first and second eigenvalue must pass the Scree-test. An alternative formulation, based on several simulated situations, could be: The ordered eigenvalues are showing an exponential decaying pattern.

If the condition, expressed as stopping rule, is fulfilled the procedure stops. If, however, this condition is not fulfilled the procedure continues. This continuation consists in applying the splitting-up technique consecutively to those time series showing positive factor loadings and those series showing negative factor loadings for the second component.

The results obtained for this recursive alternative approach are summarised in a schematic form in Figure 8.3.

FIGURE 8.3 : SPLITTING UP SCHEME




About these results at least a few important remarks can be formulated. The first remark is about the comparison of the results obtained for the recursive approach with the results obtained by using the two and three components model. It turns out that the partitioning of the whole set of time series, that could be derived from using a two and three components model, is exactly the same as that obtained at the first and second level of the recursion.

A second remark is about the behaviour of the individual time series during the whole recursive procedure. It is interesting to note that, with the exception of series 38, all the time series belonging to the same model structure stayed together during the whole recursive process.

Most of the conclusions that can be drawn from the results of the Cluster analysis are a confirmation of the conclusions already formulated in the first experiment. From the results obtained for the Centroid method, reported in Figure 8.4, it follows that the AR model structures, i.e. models 1, 2, 3 & 4, can be identified as homogeneous clusters. The MA model structures, i.e. models 5, 6, 7 & 8, seem to be more sensitive to model coefficients, resulting in more heterogeneous clusters.

FIGURE 8.4 : DENDROGRAM CENTROID METHOD

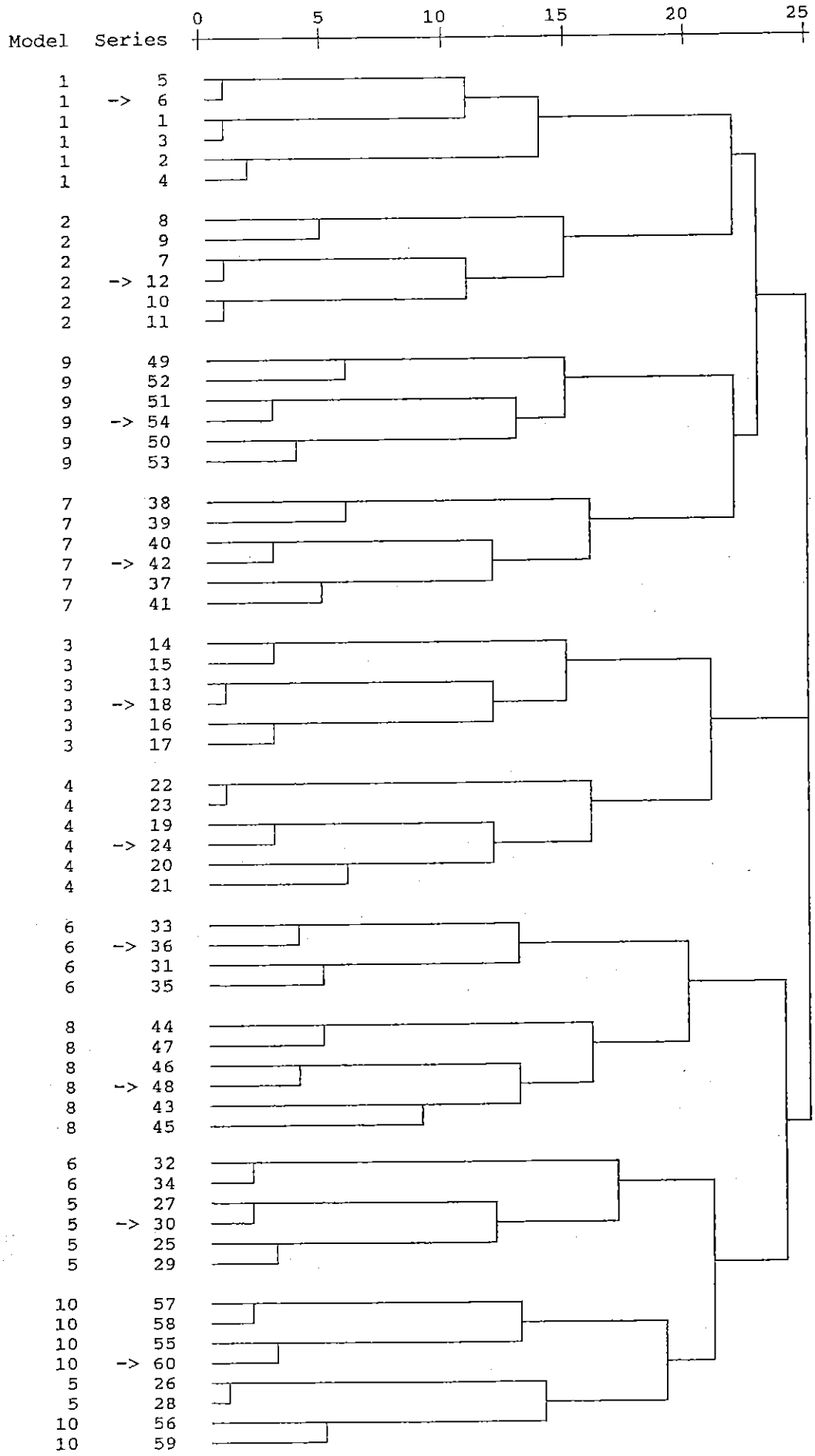
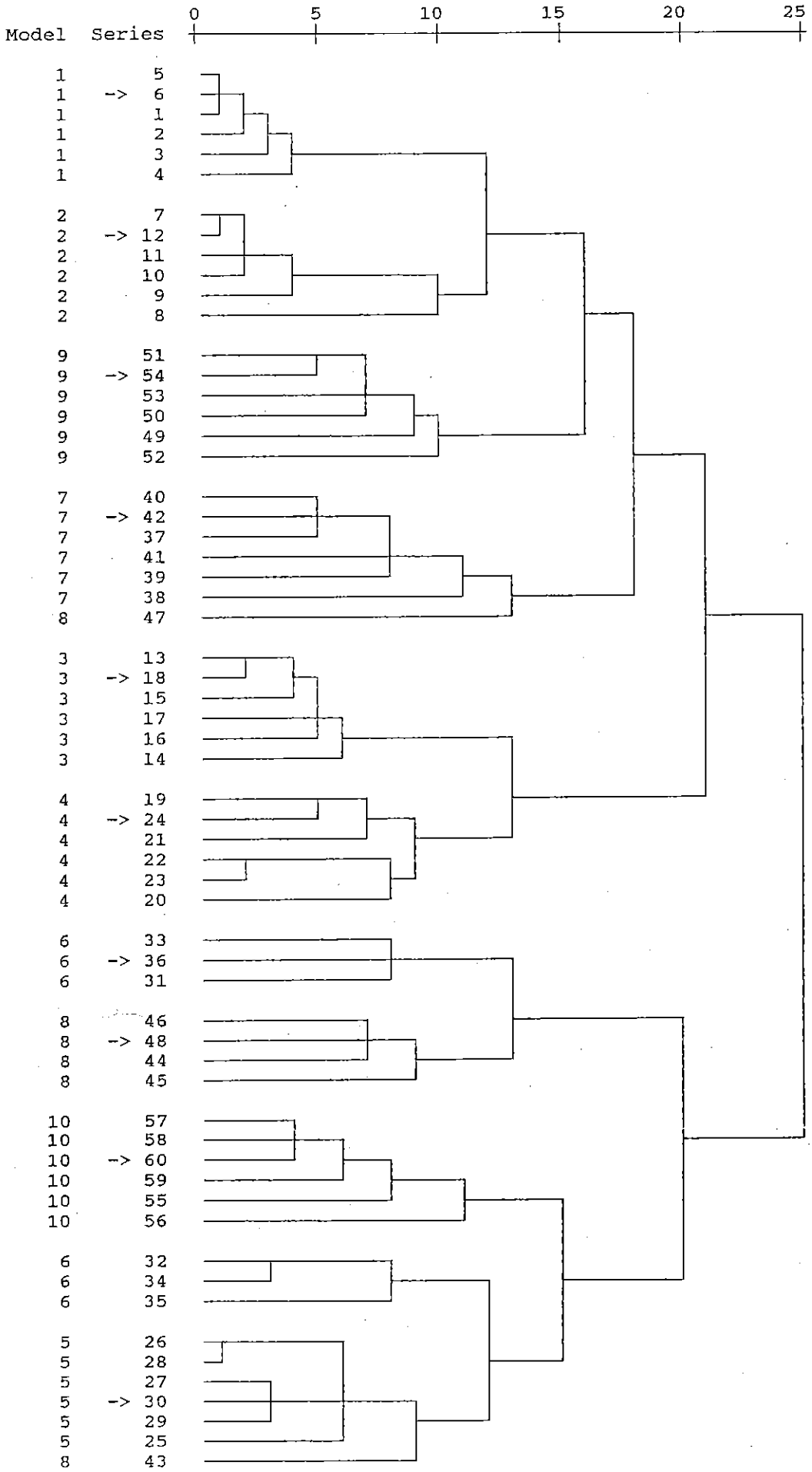


FIGURE 8.5 : DENDROGRAM AVERAGE LINKAGE METHOD



The same conclusions are also valid for the Average Linkage method, for which the results are given in Figure 8.5, i.e. homogeneous AR clusters and less homogeneous MA clusters.
Although the difference is less pronounced than in the previous experiment, the Centroid method seems to be more robust for model coefficients than the Average Linkage method.

Perhaps the most important remark that can be formulated about the results for both clustering techniques is that the composition and the hierarchical structure of the clusters are completely different from the partitioning schemes obtained for both the two and three components model as well as for the recursive approach.

## 8.3. Experiment 3

The question that will be asked in this third experiment is : what will happen if the described splitting-up technique is continued although the (remaining) time series are characterised by one and the same model structure ?

For this purpose the following situation was simulated:

| | |
|---|---|
| Number of time series | : 10 |
| Model structure | : AR(1) |
| Model coefficient | : 0.70 |
| Number of observations | : 200 |
| Residual variance | : 1 |

The resulting augmented similarity matrix, with the theoretical Π-coefficient mentioned as the 11th series, is given in TABLE 8.7 . The most striking characteristic of this similarity matrix is the pattern of the similarities obtained for the fifth series.

TABLE 8.7 : SIMILARITY MATRIX (°)

| Series | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 100 | | | | | | | | | | |
| 2 | 91 | 100 | | | | | | | | | |
| 3 | 95 | 96 | 100 | | | | | | | | |
| 4 | 85 | 95 | 90 | 100 | | | | | | | |
| 5 | 17 | 19 | 18 | 20 | 100 | | | | | | |
| 6 | 64 | 55 | 60 | 50 | 0 | 100 | | | | | |
| 7 | 98 | 89 | 93 | 84 | 16 | 66 | 100 | | | | |
| 8 | 81 | 90 | 86 | 96 | 21 | 46 | 79 | 100 | | | |
| 9 | 88 | 79 | 83 | 73 | 12 | 77 | 90 | 69 | 100 | | |
| 10 | 100 | 91 | 96 | 86 | 17 | 64 | 98 | 82 | 88 | 100 | |
| ->11 | 97 | 87 | 92 | 82 | 16 | 68 | 98 | 78 | 91 | 96 | 100 |

(°) Similarities multiplied by 100 and rounded.

TABLE 8.8 : EIGENVALUES AND ( CUMULATIVE ) PERCENTAGES

| | Eigenvalue | Percentage | Cumul.Perc. |
|---|---|---|---|
| 1 | 8.5800 | 77.9999 | 77.9999 |
| 2 | 1.0464 | 9.5129 | 87.5128 |
| 3 | .7126 | 6.4779 | 93.9906 |
| 4 | .3337 | 3.0337 | 97.0244 |
| 5 | .1282 | 1.1654 | 98.1898 |
| 6 | .0837 | .7605 | 98.9503 |
| 7 | .0417 | .3790 | 99.3292 |
| 8 | .0328 | .2985 | 99.6277 |
| 9 | .0261 | .2369 | 99.8646 |
| 10 | .0118 | .1069 | 99.9715 |
| 11 | .0031 | .0285 | 100.0000 |

TABLE 8.9 : PRINCIPAL COMPONENTS (°)

| Series | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 98 | - 3 | 2 |
| 2 | 95 | 9 | -19 |
| 3 | 97 | 4 | -10 |
| 4 | 92 | 15 | -28 |
| 5 | 19 | 89 | 41 |
| 6 | 69 | -39 | 50 |
| 7 | 97 | - 5 | 6 |
| 8 | 88 | 18 | -32 |
| 9 | 90 | -17 | 25 |
| 10 | 98 | - 2 | 1 |
| -> 11 | 97 | - 7 | 9 |

(°) Elements multiplied by 100 and rounded.

From the eigenvalues, given in TABLE 8.8 , and based on the Scree-test it follows that, strictly speaking, at most two principal components have to be extracted. The factor loadings, corresponding to the three largest eigenvalues, are given in TABLE 8.9. Using the first and second principal component the following partition of the 11 cases can be derived:

GROUP 1 : Series 1, 6, 7, 9, 10, 11
GROUP 2 : Series 2, 3, 4, 5, 8

Using the second and third principal component the partition of the 11 cases gives the following results:

GROUP 1 : Series 1, 6, 7, 9, 10, 11
GROUP 2 : Series 2, 3, 4, 8
GROUP 3 : Series 5

It follows that in the three component case the grouping remains almost the same. The only difference is that the fifth series is forming a separate third group.

In order to get a better insight into the partitioning mechanism the results from the grouping were compared with the estimated Π-coefficients. For each of the simulated series an AR(1) model was estimated. The estimation results were then confronted with the results of the partitioning. These combined results are presented in TABLE 8.10 .

Before any interpretation of the results of TABLE 8.10 can be made, it must be mentioned that, following the AIC criterion, the fifth simulated series, although simulated as an AR(1) series with coefficient 0.7, was identified as an AR(3) model. This is, without any doubt, the reason for the rather particular results obtained for the fifth series.

TABLE 8.10 : GROUPING THE Π-COEFFICIENTS

| GROUP 1 | | GROUP 2 | |
|---|---|---|---|
| Series | Coefficients | Series | Coefficients |
| 8 | 0.634 | 10 | 0.689 |
| 4 | 0.647 | 1 | 0.690 |
| 2 | 0.663 | 7 | 0.695 |
| 3 | 0.676 | -> 11 | 0.700 |
| | | 9 | 0.726 |
| | | 6 | 0.796 |
| | | 5 | 0.615 |
| | | | 0.154 |
| | | | - 0.188 |

The main conclusion that can be drawn from the results of TABLE 8.10 is that the grouping, resulting from the splitting-up technique, is based on the ( absolute ) value of the Π-coefficients. Stated in more general terms one can conclude that, given a set of time series following the same model structure, the splitting-up technique goes in the direction of discriminating within the same model structure between models with different values for their coefficients.

In order to get even a better insight into the interactions of the similarity matrix, a Cluster Analysis was carried out. The results for both the Centroid and the Average Linkage method are presented respectively in FIGURE 8.6 and 8.7.

FIGURE 8.6 : DENDROGRAM CENTROID METHOD

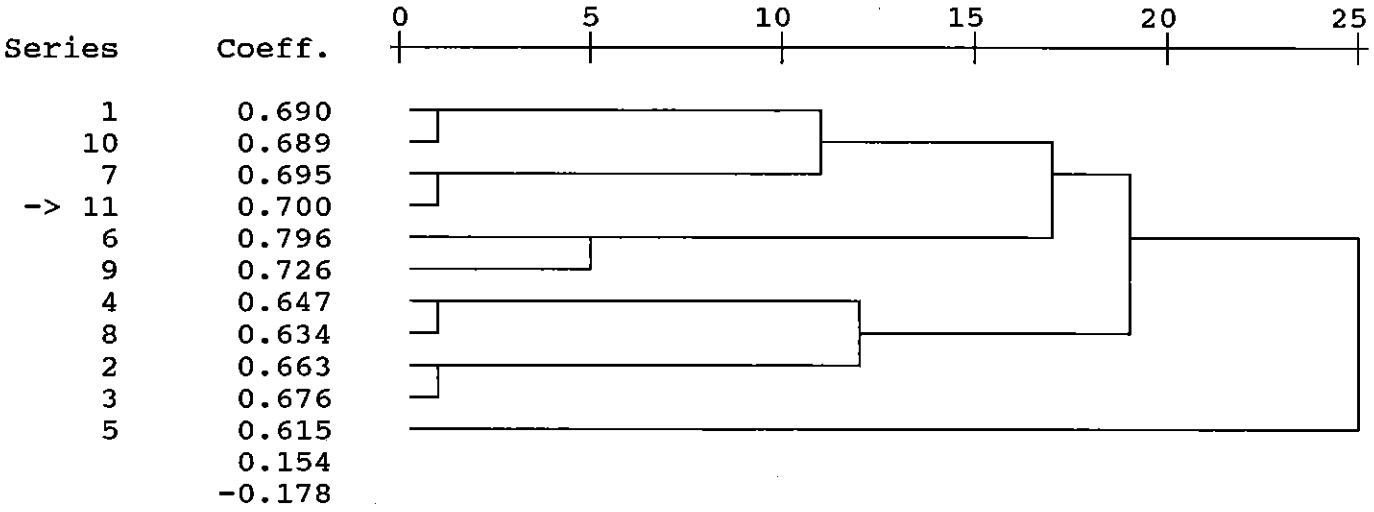
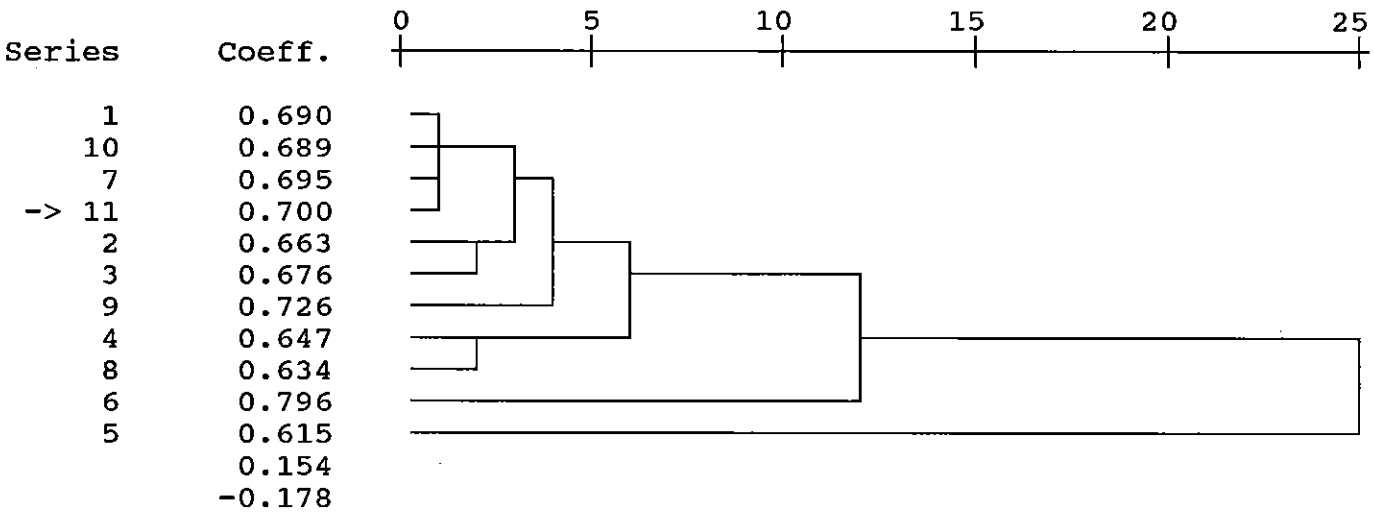



FIGURE 8.7 : DENDROGRAM AVERAGE LINKAGE METHOD




Both clustering methods are showing one common characteristic. In each of these methods the fifth series is treated as an isolated and separate cluster consisting of just one single series. This result is completely in concordance with the configuration of the three component model and the other remarks formulated about this simulated series.

However, both clustering methods are also showing a remarkable difference. Whereas the sixth series ( $\Pi = 0.796$ ! ) is treated as a single-series-cluster by the Average Linkage method, this series is linked together with the ninth series ( $\Pi = 0.726$ ) by the Centroid method.

The main conclusion of this third experiment can be summarised as follows: if the (remaining) time series are characterised by one and the same model structure, using the proposed splitting-up technique will result in discriminating within the same model structure between models with different values for their coefficients. The method seems to be much more sensitive to a difference in model structure than to a difference in the values of the model coefficients.

## 9. Conclusions

The main purpose of this paper was to investigate the possibility to identify, within a large set of weak stationary time series, subsets of time series characterised by the same model structure but for which the model coefficients may vary.

Starting from the AR-representation of weak stationary time series and by using the Euclidean distance measure, a similarity matrix was derived. Applying a Principal Component analysis on this similarity matrix led to the following conclusions:

- Since the Principal Component analysis is a deterministictechnique, a criterion to discriminate among the coefficients of the components cannot be used.

- A more robust use of these components consists in using only the sign of the factor loadings.

- By only using the sign of the factor loadings for the first three principal components, the number of subsets that can be identified is restricted to four. These four groups can only be identified under the assumption that the third eigenvalue can be considered as being "significant". ( Experiment 1 )

- An alternative recursive heuristic approach seems to succeed in the identification of homogeneous subsets, characterised by the same model structure. (Experiment 2 )

- The risk that the recursive algorithm goes one ( or even more ) level(s) to deep results in a classification that will not only be based on the model structure but also on the coefficients of the model. ( Experiment 3 )

- Since this alternative method is only based on the sign of the factor loadings, the final results will be less sensitive to the model coefficients than the results derived from a Cluster analysis.

- In order to identify homogeneous clusters none of the clustering techniques used seems to be a valuable alternative to the proposed recursive use of the Principal Component analysis.

The main advantage of the proposed method is that it is based on the Principal Component analysis, i.e. on a mathematical and deterministic technique which does not require to specify any underlying statistical model. More in particular, no assumptions have to be made about the probability distribution of the original data, i.e. the Π-coefficients of the autoregressive polynomials.

Another advantage of the recursive splitting-up method is its simplicity and the absence of any subjective intervention or evaluation. It is mainly for these reasons that the method proposed here can easily be automated. Algorithms for order determination, estimation of the Π-polynomials and calculation of the eigenvalues and vectors are generally available.

Apart from these major advantages also a few minor and less spectacular advantages must be mentioned. The facility to include in the analysis the theoretical and known AR-representations of commonly used model structures seems to be an invaluable help in the correct interpretation and identification of the different groups and clusters.

Even the fact that not only the estimated AR-polynomials may differ in length but that also the time series themselves need not be of the same length can be seen as an additional advantage.

The proposed method is also highly flexible. Distance measure, order determination and estimation method can easily be changed. By using only the seasonal lags of the Π-polynomials, the method can be used for the specific analysis of seasonal time series or to discriminate between seasonal and nonseasonal models.

All these positive arguments do not mean that all problems are solved. A lot of questions remain. In the first place, a proper formal proof would be highly desirable. Also a few other items are waiting for an answer and/or a solution. A few of these questions that are still open are:

- What about the use of nonstationary time series?
- What is the influence of the chosen measure for distance and the order criterion used?
- What is the sensitivity of the proposed method for model coefficients, seasonal model components and model factors with complex roots?

For all these reasons the only thing that can be said at this moment is that the recursive use of the Principal Components technique looks very promising.

**REFERENCES - BIBLIOGRAPHY**


[ 1] AKAIKE, H.
A New Look at the Statistical Model Identification, IEEE Transactions on Automatic Control, Ac-19, 1974, pp. 716-723.

[ 2] AKAIKE, H.
A Bayesian Extension of the Minimum AIC Procedure of Autoregressive Model Fitting, Biometrika, vol. 66, no. 2, 1979, pp.237-242.

[ 3] ANG, J. , CHUAA, J. H. & FATEMI, A. M.
How to Use Akaike Criterion to Automatically Specify and Construct Box-Jenkins Models, Journal of Business Forecasting, 1982, Spring, pp. 38-42.

[ 4] BHANSALI, R.J.
Asymptotically Efficient Selection of the Order by the Criterion Autoregressive Transfer Function, The Annals of Statistics, vol. 14, 1986, pp. 315-325.

[ 5] BORGHERS, E. & McLEOD, G.
AUTOMOD : An Application for the Belgian Car Market, in : Third International Symposium on Forecasting, Philadelphia, June 5-8, 1983, p. 111.

[ 6] BOX, G.E. & JENKINS, G.M.
Time Series Analysis : Forecasting and Control, Holden-Day, San Francisco, 1976.

[ 7] CATTELL, R.B.
The Meaning and Strategic Use of Factor Analysis, in : CATTELL, R.B. (Ed.), Handbook of Multivariate Experimental Psychology, Rand McNally, Chicago, 1966.

[ 8] CHATFIELD, C. & COLLINS, A.J.
Introduction to Multivariate Analysis, Chapman and Hall, London, 1980.

[ 9] CORDUAS, M.
Distanza tra Modelli : Problemi Metodologici e Indici Statistici, Statistica, vol. 44, no. 3, 1984, pp. 513-524.

[10] EVERITT, B.S.
Unresolved Problems in Cluster Analysis, Biometrics, vol. 35, 1979, pp. 169-182.

[11] EVERITT, B.S.
Cluster Analysis, Heinemann Educational Books Ltd., London,1980.

[12] HARKSTIAN, A.R. , ROGERS, W.D. & CATTELL, R.B.
The Behavior of Numbers of Factor Rules with Simulated Data, Multivariate Behavioral Research, vol. 17, 1982, pp. 193-219.

[13] HARMAN, H.H.
Modern Factor Analysis, University of Chicago Press, Chicago, 1976.

[14] HARTIGAN, J.A.
Clustering Algorithms, John Wiley & Sons Inc., New York, 1975.

[15] HILL, G. & FILDES, R.
The Accuracy of Extrapolation Methods : An Automatic Box-Jenkins Package SIFT, Journal of Forecasting, vol. 3, 1984, pp. 319-324.
[16] HILL, G. & WOODWORTH, D.
Automatic Box-Jenkins Forecasting, Journal of the Operational Research Society, vol. 31, no. 5, 1980, pp. 413-422.
[17] JACKSON, J.E.
A User's Guide to Principal Components, John Wiley & Sons Inc., New York, 1991.
[18] JARDINE, N. & SIBSON, R.
Mathematical Taxonomy, John Wiley & Sons Inc., New York, 1971.
[19] JENKINS, G.M.
Some Practical Aspects of Forecasting in Organizations, Journal of Forecasting, vol. 1, 1982, pp. 3-21.
[20] JENKINS, G.M. & McLEOD, G. ( Eds. )
Case Studies in Time Series Analysis - Volume 1, Gwilym Jenkins & Partners Ltd., Lancaster, 1983, pp. 113-149.
[21] KAISER, H.F.
The Application of Electronic Computers to Factor Analysis, Educational and Psychological Measurement, vol. 20, 1960, pp. 141-151.
[22] LAWLEY, D.N. & MAXWELL, A.E.
Factor Analysis as a Statistical Method, Butterworths, London, 1971.
[23] MILLIGAN, G.W.
An Examination on the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms, Psychometrika, vol. 45, 1980, pp. 325-342.
[24] MORIARTY, M. & ADAMS, A.
Issues in Sales Territory Modelling and Forecasting Using Box-Jenkins Analysis, Journal of Marketing Research, vol. XVI, May, 1979, pp. 221-232.
[25] MULAIK, S. A.
The Foundations of Factor Analysis, McGraw-Hill, New York, 1972.
[26] NEWTON, H.J.
TIMESLAB : A Time Series Analysis Laboratory, Wadsworth & Brooks / Cole Publishing Company, Pacific Grove, 1988.
[27] PARZEN, E.
Multiple Time Series : Determining the Order of Approximating Auto-regressive Schemes, in : KRISHNAIAH, P. ( Ed. ), Multivariate Analysis IV, North-Holland, Amsterdam, 1977, pp. 283-295.
[28] PICCOLO, D.
A Distance Measure for Classifying ARIMA Models, Journal of Time Series Analysis, vol. 11, no. 2, 1990, pp. 153-164.

[29] POULOS, L. , KVANLI, A. & PAVUR, R.
A Comparison of the Accuracy of the Box-Jenkins Method with other Automated Forecasting Methods, International Journal of Forecasting, vol. 3, 1987, pp. 261-268.

[30] REILLY, D.P.
Recent Experiences with an Automatic Box-Jenkins Modelling Algorithm, in: ANDERSON, O.D. & PERRYMAN, M.R. ( Eds. ), Time Series Analysis, North-Holland, Amsterdam, 1981, pp. 493-508.

[31] SHIBATA, R.
Various Model Selection Techniques in Time Series Analysis, in: HANNAN, E. J. , KRISHNAIAH, P. R & RAO, M. M. ( Eds. ), Handbook of Statistics, Volume 5: Time Series in the Time Domain, North-Holland, 1985, pp. 179-187.

[32] SNEATH, P.H.A. & SOKAL, R.R.
Numerical Taxonomy, W.H. Freeman & Co, San Francisco, 1973.

[33] TIAO, G.C. & TSAY, R.S.
Consistency Properties of Least Squares Estimates of Autoregressive Parameters in ARMA Models, The Annals of Statistics, vol. 11, no. 3, 1983, pp. 856-871.

[34] TSAY, R.S. & TIAO, G.C.
Consistent Estimates of Autoregressive Parameters and Extended Sample Autocorrelation Function for Stationary and Nonstationary ARMA Models, Journal of the American Statistical Association, vol. 79. no. 385, March, 1984, pp. 84-96.

[35] ZWICK, W.R. & VELICER, W.F.
Comparison of Five Rules for Determining the Number of Components to Retain, Psychological Bulletin, vol. 99, 1986, pp. 432-442.

**REFERENCES - SOFTWARE**

[36] AUTOBOX
Automatic Forecasting Systems, Inc.
Hartboro, USA.
[37] AUTOMOD
Gwilym Jenkins & Partners Ltd.
Lancaster, England.
[38] MATHEMATICA
Wolfram Research, Inc.
Champaign, USA.
[39] SCA Statistical System - Workstation / Mainframe Version
Scientific Computing Association Corp.
Chicago, USA.
[40] S-PLUS
Statistical Sciences, Inc.
Seattle, USA.
[41] SPSS
SPSS Inc.
Chicago, USA.
[42] TIMESBOARD Computing Library
Institute of Statistics
Texas A&M University
College Station, USA.
[43] TIMSAC - 78
The Institute of Statistical Mathematics
Tokyo, Japan.

**LIJST VAN RECENTE SESO-RAPPORTEN**

**VERHETSEL A. & A. JORISSEN,** België, een land met meer dan twee snelheden ? Het gebruik van jaarrekeninggegevens in economisch-geografisch onderzoek, november 1992, 38 blz. **(92/279)**

**SCHROYEN F.,** The comparative statics of tax evasion with elastic labour supply - Can we really say anything about the reactions of a tax evader ?, November 1992, 62 blz. **(92/280)**

**COPPIETERS P. & A. HUFKENS,** The role of the service sectors within the changing economic structure, December 1992, 23 blz. **(92/281)**

**COPPIETERS P., DE BORGER B. & A. HUFKENS,** Productivity growth measurement in the Belgian banking sector 1985-1989, December 1992, 26 blz. **(92/282)**

**VAN GOMPEL J.,** Stabilization with wage indexation and exchange rate flexibility -a survey of the literature-, January 1993, 47 blz. **(93/283)**

**NONNEMAN W.,** De kosten van instellingen van hoger onderwijs buiten de universiteit, januari 1993, 33 blz. **(93/284)**

**HEYLEN F., VAN POECK A. & J. VAN GOMPEL,** Real versus nominal convergence - national labour markets and the European integration process, April 1993, 42 blz. **(93/285)**

**KESENNE S.,** The unemployment impact of a basic income, May 1993, 25 blz. **(93/286)**

**CORTENS I. en W. NONNEMAN,** Is het onderwijs meer dan een filter ?, juni 1993, 31 blz. **(93/287)**

**VAN TRIER W.,** James Meade and his "social dividends" - An intriguing chapter in the history of an idea, July 1993, 34 blz. **(93/288)**

**HEYLEN F.,** The theory and evidence of unemployment persistence - A review of the literature, August 1993, 41 blz. **(93/289)**

**YZEWYN D.,** Input-outputanalyse en toeristische impactmeting - een verkennende toepassing voor Vlaanderen, augustus 1993, 70 blz. **(93/290)**

**HEYLEN F. en A. VAN POECK,** Government preferences and equilibrium inflation - a simple test of the Barro-Gordon Model, August 1993, 24 blz. **(93/291)**

**VAN POECK A.,** Belgian banks under stress - the effects of deregulation and financial innovation on the performance of the Belgian credit institutions, September 1993, 32 blz. **(93/292)**

**VAN GOMPEL J.,** Unemployment and monetary integration - on the relevance of wage formation characteristics in a target zone regime versus an irrevocably-fixed exchange rate regime, November 1993, 22 blz. **(93/293)**