Clustering techniques for the detection of Business Cycles

W. Theis[1], C. Weihs[2]

[1] SFB 475 "Komplexitätsreduktion in multivariaten Datenstrukturen"
Universität Dortmund, D–44221 Dortmund
e-mail: theis@amadeus.statistik.uni-dortmund.de

[2] Lehrstuhl für Computergestützte Statistik
Universität Dortmund, D–44221 Dortmund
e-mail: weihs@gigamain.statistik.uni-dortmund.de

Abstract

In this paper business cycles are considered as a multivariate phenomenon and not as a univariate one determined e.g. by the GNP. The subject is to look for the number of phases of a business cycle, which can be motivated by the number of clusters in a given dataset of macro-economic variables. Different approaches to distances in the data are tried in a fuzzy cluster analysis to pursue this goal.

KEY WORDS: Business cycles, cluster analysis, fuzzy clustering

1 Introduction

In the economic literature business cycles are often considered as a univariate time series phenomenon. For example business cycle phases are defined by an increase or decrease, resp., of the growth rate of the GNP. Instead, it should be recognized as a multivariate problem, which is influenced by the interplay of different economic variables. Diebold and Rudebusch (1996) discuss two main aspects of the old definition of Burns and Mitchell (1946) of the business cycle: the comovement of important economic variables and the partition of the cycle in different phases, which are assigned to different economic regimes. The proposal of a partition into different economic regimes leads to the idea that such regimes should be identifiable by some clustering algorithm.

Another reason to use clustering in this framework, was that there are a lot of different proposals how many different phases make up a business cycle. Most commonly two phases called upswing and downswing are considered sufficient. But also three to eight different phases are discussed in the literature (cp. Tichy (1976)). Heilemann and Münch (1996) discuss a 4–phase–scheme which consists of the phases upswing, upper turning point phase, downswing and lower turning point phase. For convenience the turning point phases will be called only upper and lower turning points in the rest of the paper. This classification will be compared to our clustering results.

In the next chapter we will discuss general problems of clustering in sets of economic data and present first results. We propose a new "distance" to distinguish between directions of change but discard it because of its asymmetry. Instead we normalize the data with the euclidean norm, which is shown to be appropriate for our problem. After this we discuss briefly why the usual standardization is not appropriate in this framework. Then we offer a possible interpretation of the found clusters and answer the question posed in the beginning: How many different economic regimes can be found by empirical means?

2 Problems with economic variables

The data set consists of 13 so called stylized facts for the german business cycle listed in table 1 and 157 quarterly observations from 1955/4 to 1994/4 (price index base=1991, y=yearly growth rates).

Abbr.	variable
Y	GNP, real (y)
C	Private consumption, real (y)
GD	Government deficit, percent of GNP
L	Wage and salary earners (y)
X	Net exports, percent of GNP
M1	Money supply M1 (y)
IE	Investment in equipment, real (y)
IC	Investment in construction, real (y)
LC	Unit labour cost (y)
PY	GNP price deflator (y)
PC	Consumer price index (y)
RS	Short term interest rate, nominal
RL	Long term interest rate, real

Table 1: The 13 Stylized Facts

These 13 variables have been selected by Heilemann and Münch from a total of 120 variables.

Figure 1 shows simultaneous boxplots for GNP (Y), Wage and Salary Earners (L), Investment in Equipment (IE) and Government Deficit (GD) divided into the four phases. The first three of them are variables for which the values can be well divided into groups and these groups are related to business cycle phases as will be seen in subsection 2. The last, GD, is very badly separated. "1" denotes upswing, "2" upper turning point, "3" downswing and "4" lower turning point. It is not surprising, that upswing and upper turning point have large overlaps as well as downswing and lower turning point. But the complete overlap in the boxplots of downswing and upswing even in the "well-behaved" stylized facts implies that even those two phases might not be well separated. So *well separated groups* should not be expected. How different the stylized facts behave with respect to the separation into phases is discussed in greater detail in Theis, Vogtländer, Weihs (1999).

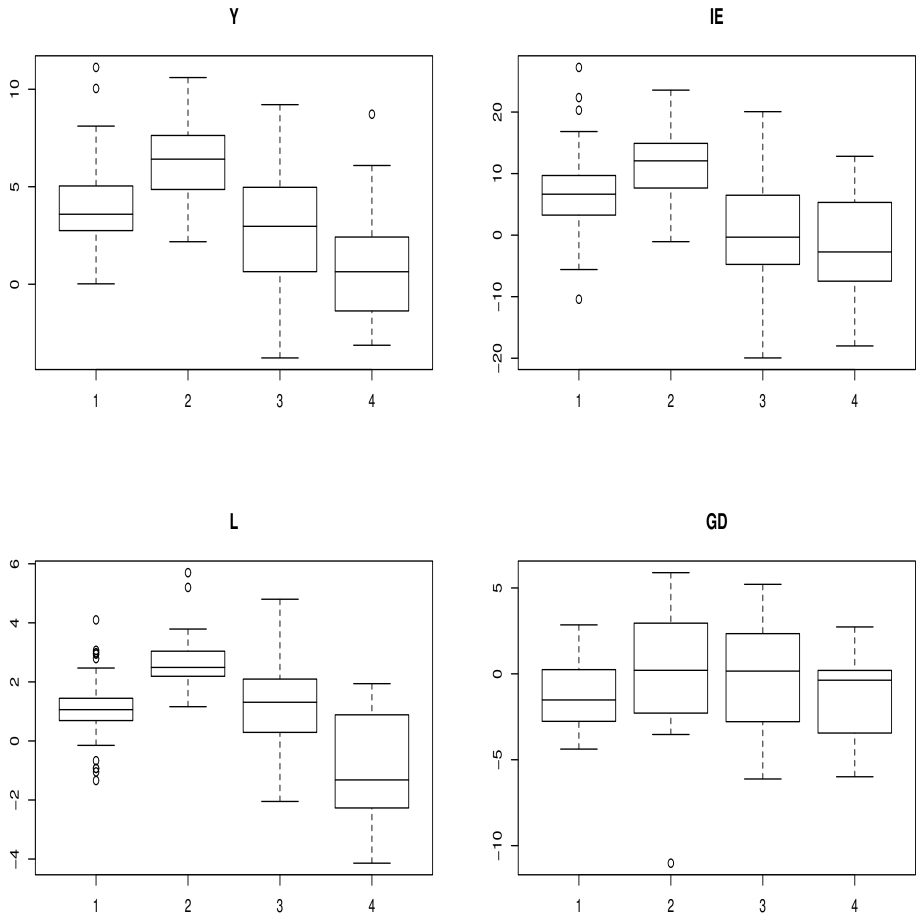

Figure 1: Boxplots of Stylized Facts within phases, Y $\hat{=}$ GNP, L $\hat{=}$ Wage and Salary Earners, IE $\hat{=}$ Investment in Equipment and GD $\hat{=}$ Goverment Deficit

Thus, usual clustering algorithms such as k-means clustering will not lead to appropriate clusters because they search for well separated groups. Instead, we used a fuzzy version of k–means.

2.1 Fuzzy–Clustering

Fuzzy–Clustering does not divide the data into well–separated groups but gives every point a probability (membership) to belong to a certain group. Figure 2 shows on the left hand side which sort of sets k–means–clustering

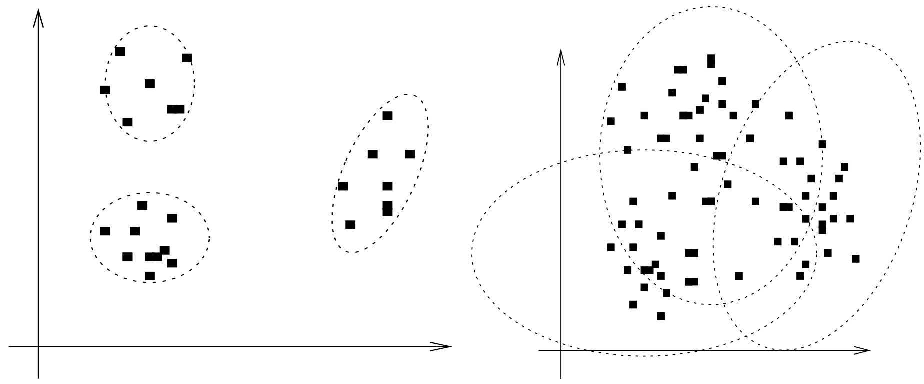

Figure 2: Difference between hard partition and fuzzy–partition

is looking for and on the right hand side a typical data set for the fuzzy approach and the sort of groups constructed by it.
Points x_i in overlapping regions get memberships u_{iv} smaller than 1 to belong to a specific group v, whereas points lying in only one group get a membership of 1 to belong to this group and 0 for all other groups. So it is easily seen, that the fuzzy–partition coincides with the hard clustering if there are well separated groups of data points.
The memberships of the n data points in a data set are summarized in a so–called membership matrix, where k denotes a given number of clusters:

$$U := (u_{iv})_{i=1,\ldots,n;v=1,\ldots,k}\,.$$

We use fuzzy-k-mean clustering as implemented in the R/S–function FANNY (Kaufman, Rousseeuw (1992)). This function minimizes the following term:

$$\sum_{v=1}^{k} \frac{\sum_{i,j=1}^{n} u_{iv}^2 u_{jv}^2 d^2(x_i, x_j)}{2\sum_{i=1}^{n} u_{iv}^2}\,,$$

here $d(x, y)$ denotes a distance measure.
To evaluate a fuzzy–partition, a measure for the goodness of "separation" into groups is needed. Such a measure is the Dunn-coefficient

$$F_k(U) := \sum_{v=1}^{k} \sum_{i=1}^{n} \frac{u_{iv}^2}{n} \quad .$$

It lies between $\dfrac{1}{k}$ for no partition (total fuzziness) and 1 for a hard partition. By normalizing to $[0, 1]$ one gets

$$\tilde{F}_k(U) = \frac{kF_k(U) - 1}{k - 1}\,.$$

For the rating of the results we use this last coefficient.

2.2 First Results

Applying FANNY directly to the whole data set and assuming two clusters results in total fuzzyness as can be seen in table 2. But from the study of parallel boxplots we know that some of the Stylized Facts have completely overlapping ranges in the different phases. So we constructed a greedy search algorithm to find the best combination of Stylized Facts to separate between groups. This algorithm deletes each variable once and applies FANNY to the new data set. The variable for which the Dunn–coefficient is increased the most in this step, is deleted from the data set for the rest of the algorithm and the search is applied to the resulting data set. Doing so we get the results listed in table 2.

Stylized Facts	$\tilde{F}_D(U)$
All	4.6629367e-15
without IC	2.597922e-14
without IC, M1	7.2164497e-14
L,IE,PY,PC	0.23714597
L, IE	0.3265420

Table 2: Normalized Dunn-Coefficient for different sets of unmanipulated Stylized Facts

The best separation is possible with Wage and Salary Earners (L) and Investment in Equipment (IE).

That this partition has something to do with business phases can be seen in figure 3 (the line represents the classification into the 4–phase–scheme, representing upswing with 0.6, upper turning point at 0.5, downswing at 0.4 and lower turning point at 0.3). The memberships in the first group found by FANNY resembles upswing combined with upper turning point.

But that the Dunn-coefficient is only 0.33 and that $\frac{1}{10}$ of the data has memberships around 0.5 indicate strongly that the data is not very well separated. This led to the idea to look for a more appropriate distance to describe the difference between business cycle phases.

Because the phases describe directions of development in the economy, a measure which reflects especially the idea of turning points is searched for. Since "turning points" should describe a change in direction from one time point to the next, we define the measure as outlined in the next section.

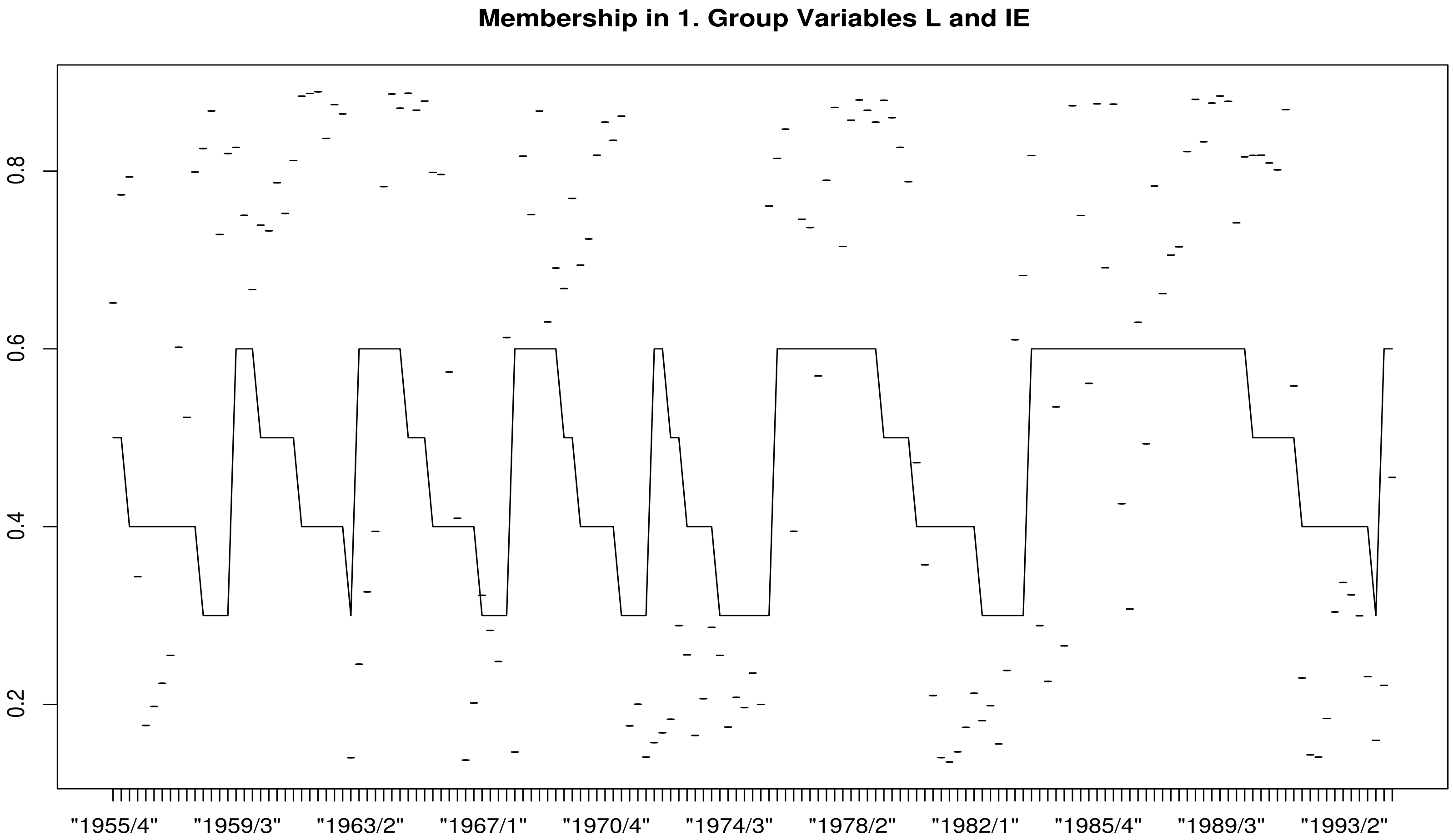

Figure 3: Memberships in 1. Group using raw data cp. to 4-phase-scheme

3 A new Distance

To distinguish between directions of development of an economy we propose the following "Distance":

Definition 1 *Let* $(X_t)_{t\in\mathbb{N}}$ *a multivariate time series in* $\mathbb{R}^d$ *and* $(x_t)_{t\in\mathbb{N}}$ *the corresponding realizations. Then define*

$$\begin{aligned}\Delta: \quad & \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R} \\ & (x_t, x_i) \mapsto \frac{1}{2}\|x_t - x_{t-1}\|\|x_i - x_{t-1}\| \cdot \\ & \sin\left(\arccos\left(\frac{< x_i - x_{t-1}, x_t - x_{t-1} >}{\|x_i - x_{t-1}\|\|x_t - x_{t-1}\|}\right)\right), \quad i \neq t\end{aligned}$$

The complicated term on the right hand side is the area of the triangle defined by the points x_t, x_{t-1}, x_i. It is a multivariate formulation of the usual formula $\frac{a \cdot b \cdot \sin\alpha}{2}$ for calculating the area of a triangle. Here the length of the basis a is $\|x_t - x_{t-1}\|$ and b is the length of another edge of the triangle, e.g. $\|x_i - x_{t-1}\|$. The arccos-term is the angle α between $x_t - x_{t-1}$ and $x_i - x_{t-1}$ because for the scalar product holds

$$< x, y >= \cos\alpha \text{ , with } \|x\| = 1 = \|y\| \,.$$

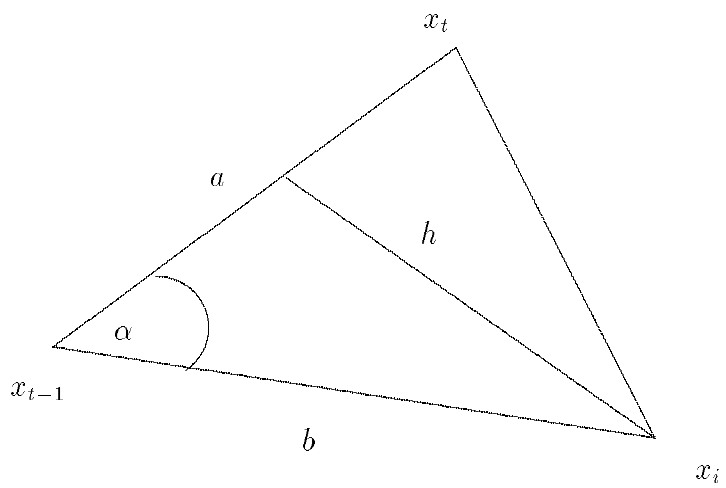

Figure 4: Calculation of the area of a triangle

The area gets smaller if x_t and x_i lie in the same direction relative to x_{t-1}. So Δ measures how similar the direction of change from x_{t-1} to x_t is to the change of the economy from x_{t-1} to x_i. Figure 5 illustrates this idea.

Applied to some part of a time series of length m (here $m = 8$) the distances $\Delta(i, j)$ build a matrix of dimensions m and $m - 1$ of the following form:

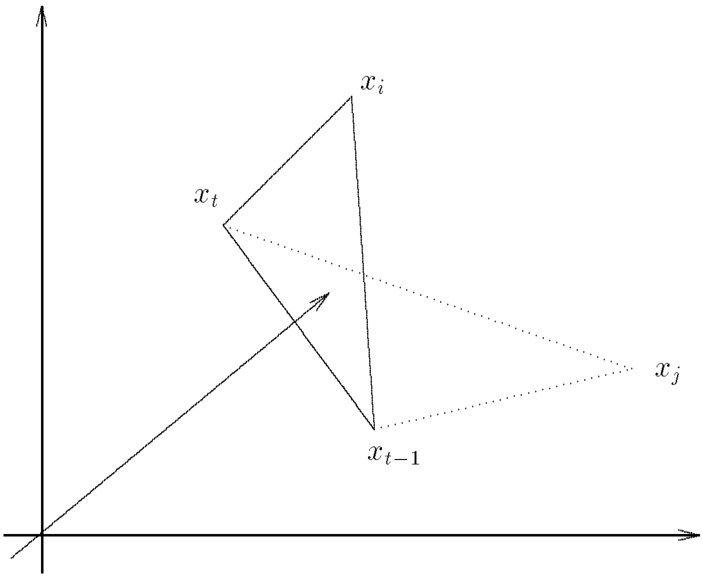

Figure 5: Illustration of the triangle-distance

	2	3	4	5	6	7	8
1	0	7.93294	22.8509	13.4588	25.4090	17.0766	13.1911
2	0	0	25.0411	11.756	21.9319	18.0507	14.9415
3	7.93294	0	0	8.7315	15.0215	**8.09695**	5.56038
4	10.1469	25.0411	0	0	6.57221	8.41444	12.0736
5	12.2092	23.7597	8.7315	0	0	5.85778	9.0408
6	15.9978	18.5606	**20.6145**	6.57221	0	0	3.59261
7	15.0567	9.32700	20.4663	8.87859	5.85778	0	0
8	12.7051	5.14122	14.1319	7.70673	10.9967	3.5926	0

So no usual clustering can be performed on this matrix because the algorithms use only the upper or lower half of distance matrices which are — due to the definition of metrics — symmetric. Symmetrization leads to the loss of the main characteristics. Thus, the "distance" measure Δ was discarded. The next section describes our alternative. Instead of manipulating the distance, we manipulate the data with the aim to reveal the direction of development.

4 Normalizing the data

4.1 Idea

Normalizing the data points with the euclidean norm reduces the information in the data points to the direction in p dimensional space and therefore the direction relative to the origin is compared by the euclidean distance

$d(i,j)$. So we now consider the distance $d(i,j) := \left\| \frac{x_i}{\|x_i\|} - \frac{x_j}{\|x_j\|} \right\|$. Figure 6 illustrates this idea.

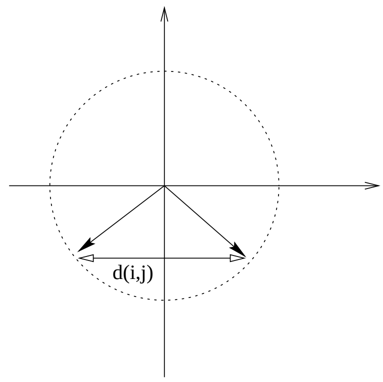

Figure 6: Illustration of distances of normalized data

4.2 Results

Table 3 lists the normalized Dunn–coefficient for the normalized data. As in the case of unnormalized data (cp. table 2) using all variables and two groups leads to a Dunn–coefficient near 0. So again a greedy search for the best subset of variables is performed where at each step the new data set is normed. Again *Wage and Salary earners (L) and Investment in Equipment (IE)* are the best Stylized Facts for the clustering. But here the Dunn-coefficient is more than twice as high as in the unnormalized case. Note that the set of the best seven variables leads to a higher $\tilde{F}_D(U)$ than four variables in the unnormalized case.

Stylized Facts	$\tilde{F}_D(U)$
All	4.440892e-16
without X	2.420286e-14
Best 8	0.20720696
Best 7	0.2571714
Y, C, L, IE	0.460536
Y, L, IE	0.549997
L, IE	0.71484

Table 3: Normalized Dunn–coefficients with nomalized Data

Figure 7 shows fuzzy–memberships in the first group for normalized data vs. the classification of Heilemann, Münch as in figure 3 (upswing $\hat{=}$ 0.6, downswing $\hat{=}$ 0.4). The superiority of the normalized clustering can be seen, compared to figure 3. The memberships are much more often nearly 1 and stay high during upswing and upper turning point.

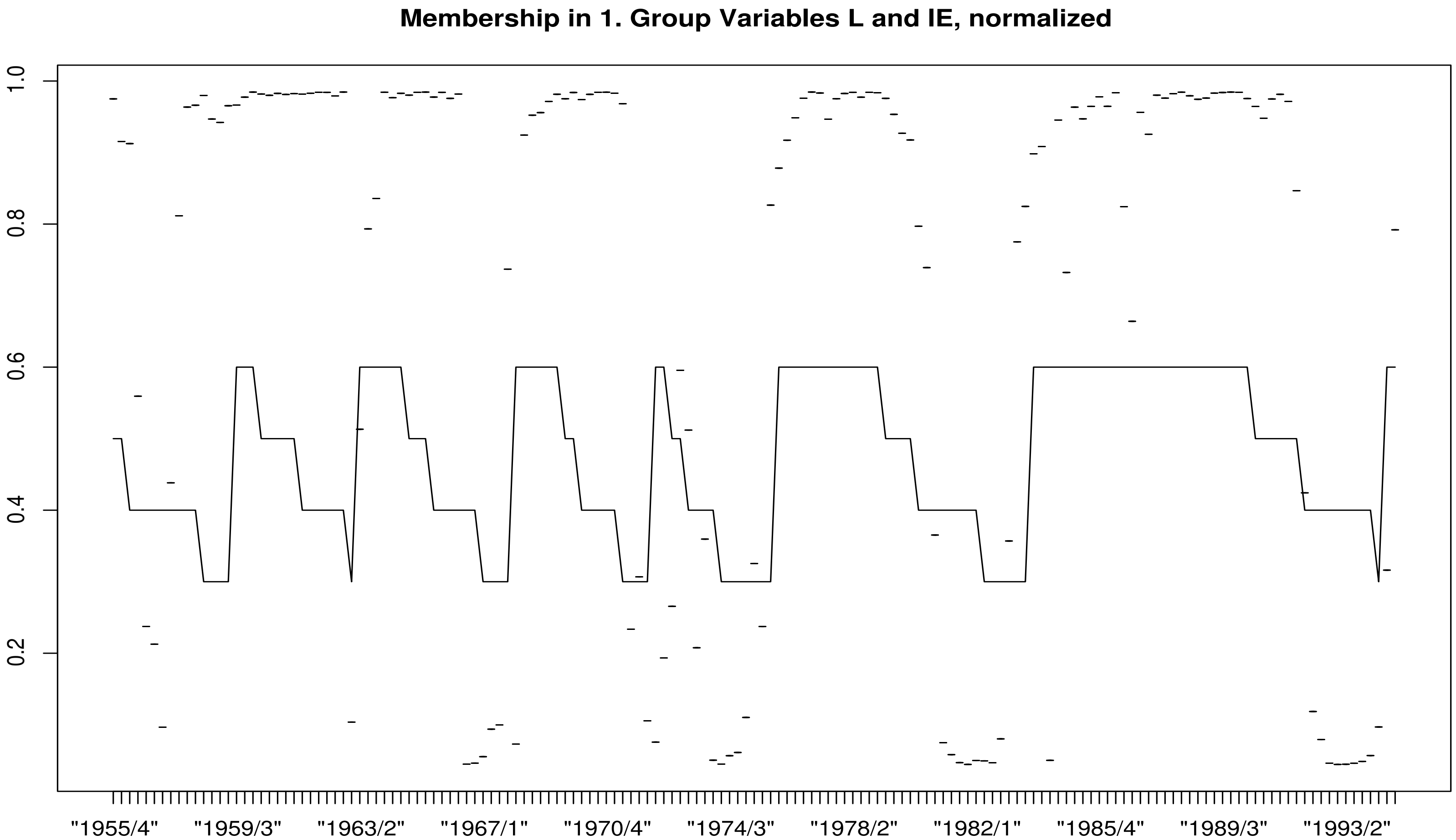

Figure 7: Memberships in group 1 for normalized data and Variables L and IE

In figure 7 a difficulty can be spotted around 1971 with memberships around 0.5 in group 1. This will be discussed in section 6.

Figure 7 suggests that high membership in the first group indicates upswing or upper turning point. The bar–chart in figure 8 strengthens this impression. On the other hand neither downswing nor lower turning point are entirely in group 2. But if one looks at figure 7 it is easy to see that most of the errors are made in the beginning of our time period, when the boom after World War II still lasted. So these "errors" are due to small differences between the main phases in this time.

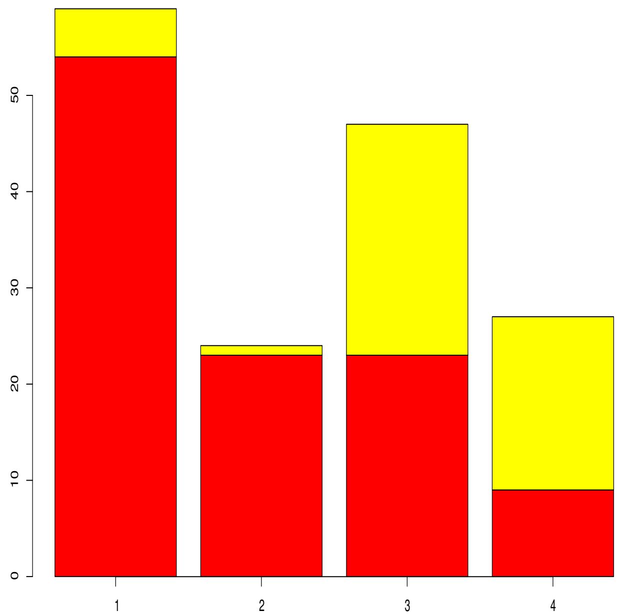

Figure 8: Bar-Chart of Membership to 4-phase-classification (Group 1 $\hat{=}$ dark)

5 Why not Standardization?

The reason not to use standardization of variables instead of normalization of observations is that standardizing by mean and standarddeviation does not support the idea to look for directions in space because it changes the components of the vectors differently according to the overall behaviour of the corresponding variables. This destroys the original directions and figure 9 shows that this leads to a clustering less related to business cycle phases than the clustering of the original data.

In our case all standarddeviations are greater than 1 (sd=(2.974, 2.756, 2.697, 1.689, 2.143, 4.512, 8.865, 7.481, 3.318, 1.674, 1.802, 2.483, 1.457)) so that the standardization leads to a concentration of the data around the origin. Due to the concentration around the origin, the differences between the datapoints decrease further which is another reason that the clustering gets even worse compared to clustering of the original data as can be seen by the normalized Dunn–coefficients in table 4.

Stylized Facts	$\tilde{F}_D(U)$
All	0.0
without X	1.776357e-15
LC, PY,PC,RS	0.1312714
LC, PY	0.2397149

Table 4: FANNY with standardized Data

6 Are there more than two groups?

The aim of using cluster analysis was to determine how many different economic regimes could be found empirically. Up to now only two groups were found. Testing different numbers of clusters showed that more than three groups can not be separated in our data set. Using four or more groups only leads to a split in the memberships in one of the larger groups and the normalized Dunn–coefficient decreases significantly with each new group.

The clustering into three groups however reveals an interesting third group. Groups no. 1 and 3 are essentially the two groups from figure 7 but group no. 2 contains the difficult time period around 1971 — known to be a time of exceptional economic behaviour caused by the first oil–crisis (see figure 10).

Table 5 reports the normalized Dunn–coefficients and selected variables in the case of clustering into three groups. Notice, that the best two variables are again Wage and Salary Earners (L) and Investment in Equipment (IE), in the selection of the best four variables GNP (Y) and Private Consumption (C) are replaced by their respective price index.

Stylized Facts	$\tilde{F}_D(U)$
All	1.04639e-14
L, IE, PY, PC	0.26673257
L, IE, PC	0.39876288
L, IE	0.63404132

Table 5: Normalized Dunn–coefficients with nomalized Data

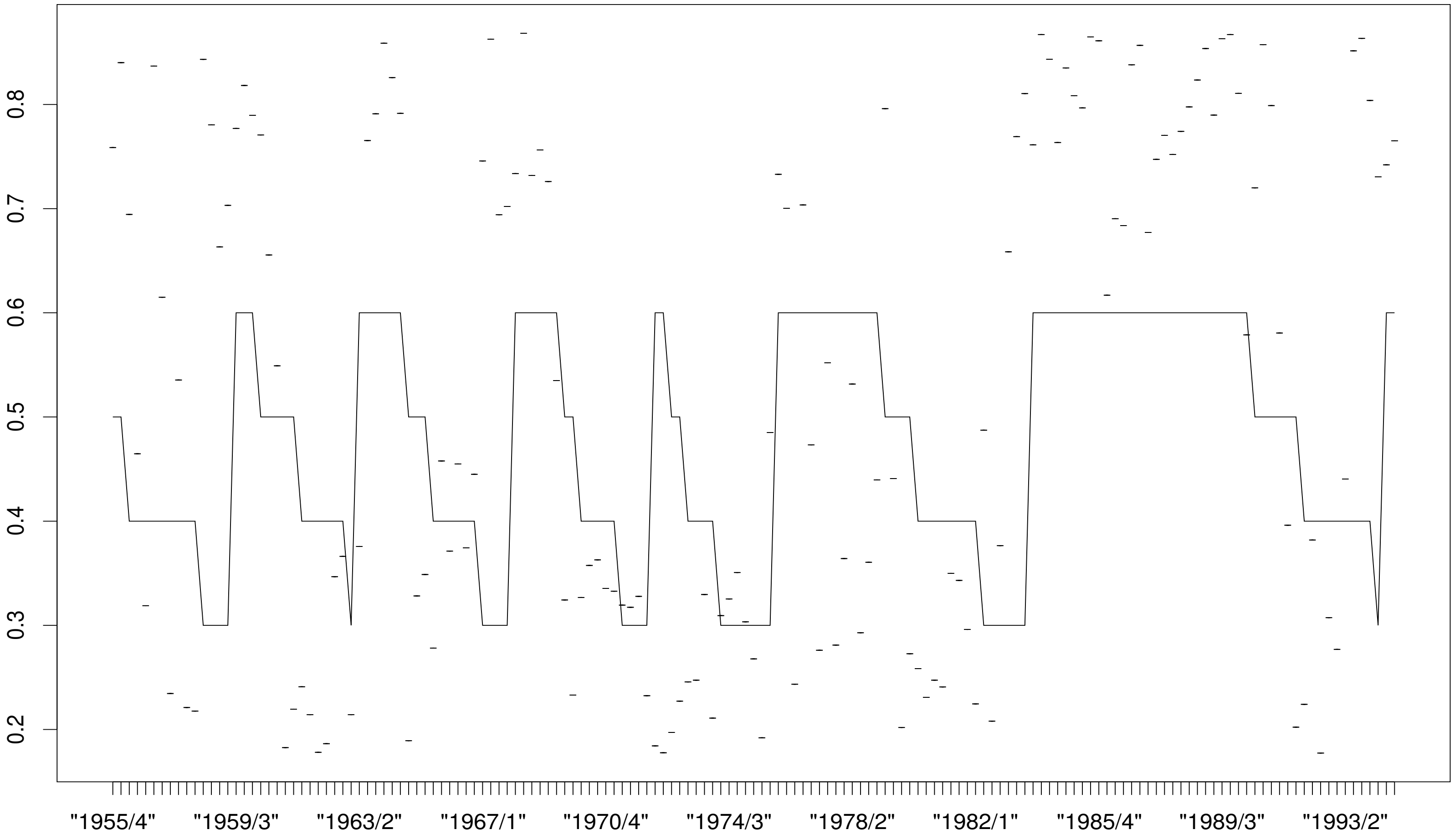

Figure 9: Clustering standardized Data and RWI-Phases

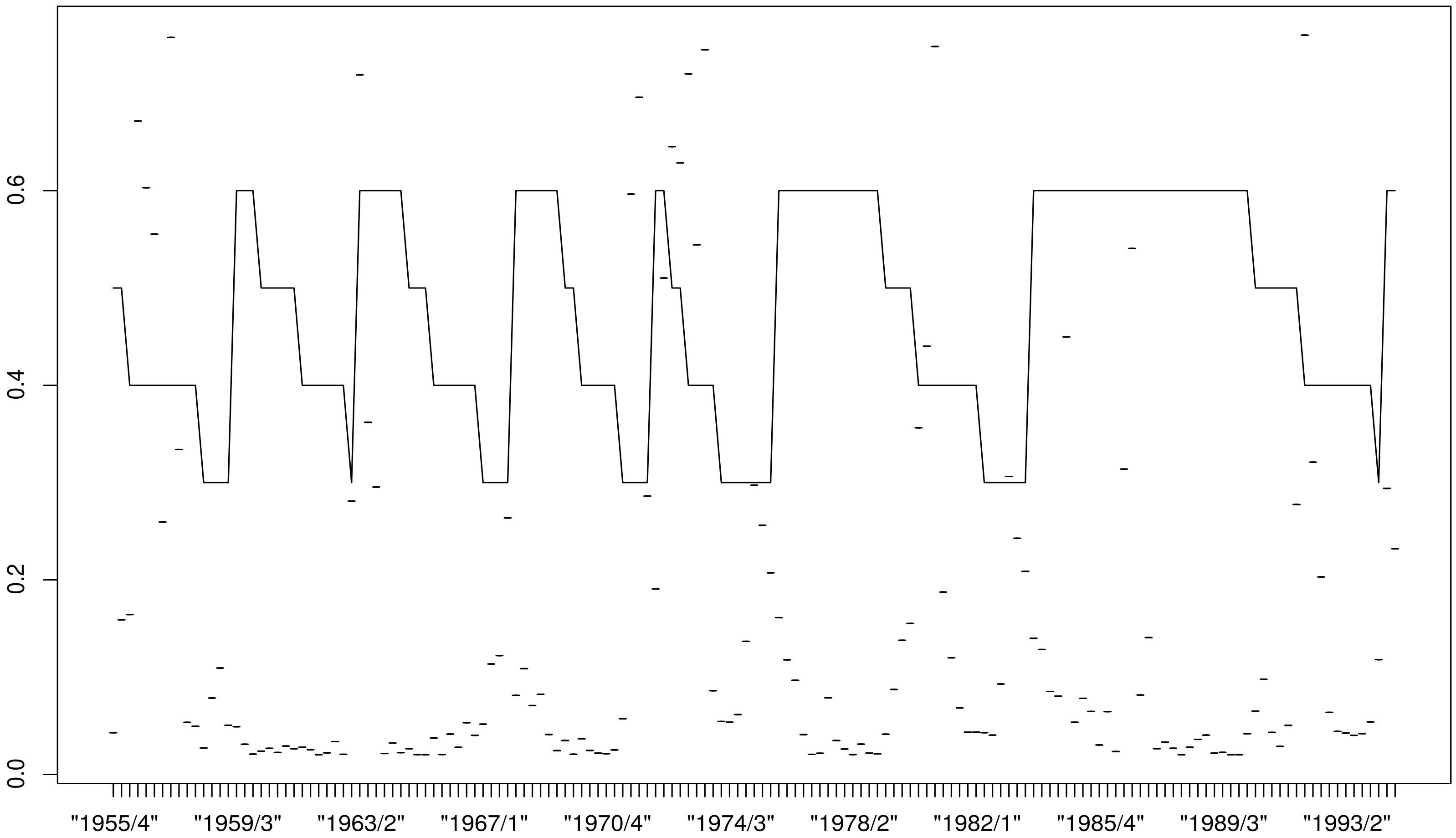

Figure 10: Membership in Group 2, using 3 groups and normalization

7 Conclusions

Using data vectors normalized by the euclidean norm and clustering thereafter has proved to be capable of finding different economic behaviour in the data. Especially the identification of exceptional economic conditions is very promising. It shows that with this method it is possible to detect unusual economic behaviour in the search for different economic regimes corresponding to business cycle phases.

The need for a reduction to a proper set of variables is a sign that some of the theoretically important variables produce too much noise and eliminate thereby the existing discernable groups.

Acknowledgment

This work has been supported by the Collaborative Research Centre "Reduction of Complexity in Multivariate Data Structures" (SFB 475) of the German Research Foundation (DFG).

References

A. F. Burns, W. C. Mitchell (1946), Measuring business cycles , NBER, New York

F. X. Diebold, G. D. Rudebusch (1996), Measuring business cycles: a modern perspective, *The Review of Economics and Statistics* **78**, 67-77

Kaufman, L. and Rousseeuw, P.J. (1990): Finding Groups in Data, 164–197 *Wiley, New York*

Heilemann, U. and Münch (1996) "'West German Business Cycles 1963-1994: A Multivariate Discriminant Analysis"', Paper presented at the 1995 CIRET-Conference in Singapore, CIRET-Studien 50, München

Tichy, G. (1976) Konjunkturschwankungen, *Heidelberger Taschenbücher, 174*, Berlin.

Theis, W., Vogtländer, K., Weihs, C. (1999) Descriptive Study of the RWI data set, Technical Report, **45/1999**, SFB 475, Universität Dortmund