Outliers and Influence Points in German Business Cycles

Manuela Zucknick, Claus Weihs*, and Ursula Garczarek

* e-mail: weihs@statistik.uni-dortmund.de

Collaborative Research Centre
and
Department of Statistics,
University Dortmund

Abstract

In this paper, we examine the German business cycle (from 1955 to 1994) in order to identify univariate and multivariate outliers as well as influence points corresponding to Linear Discriminant Analysis. The locations of the corresponding observations are compared and economically interpreted.

Keywords: outliers, influence points, linear discriminant analysis, business cycles

1 Introduction

The aim of this paper is to examine the German business cycle (from 1955 to 1994) in order to identify potential outliers. Here, we use the term "outlier" for observations either having extreme values or having a large influence on the model fit. Business cycles are multivariate time series, so we have to determine the outliers in this specific context. Often, multivariate time series are just treated as an ensemble of univariate time series. This means that the aspect of dependency between the components is not taken into account, and only univariate types of outliers are examined. But in a multivariate time series many different types of outliers – in the sense of extreme or influential observations - can exist, both, multivariate and univariate ones (for more detailed information see Tsay et al. (2000)). In a specific time series, those different types of outliers do not necessarily correspond to the same observations, since they are based on different definitions.

In this paper, we apply several simple methods to identify three kinds of extreme or influential observations. First, we determine univariate potential outliers by a simple 3σ rule. Secondly, we try to find possible multivariate outliers by fitting an AR(1) process and comparing the model fits by using the AIC's. Thirdly, we identify observations which have a large influence on the discriminant functions, when a Linear Discriminant Analysis is performed. We call these observations influential.

In the next section, we describe the data set. In sections 3 to 5, we try to detect potential univariate outliers, multivariate outliers and influential observations, separately. Then, we discuss the results and compare the three determined sets of extreme or influential observations. Finally, a short conclusion is presented.

2 The Data Set

The data are available in form of a multivariate time series consisting of 13 economic variables observed at 157 successive quarterly observations from 1955/4 to 1994/4 (price index base is 1991). The economic factors, which have been obtained to describe the German business cycle, are given in Table 1.

An experts' classification of the data into business cycle phases was obtained by Heilemann and Münch (1996). A four phase scheme is used dividing the business cycle into *upswing (1)*, *upper turning point (2)*, *downswing (3)* and *lower turning point (4)* phases. During the time period used for this data set the German business cycle passed through these phases cyclically according to the scheme $4 \rightarrow 1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 1$ etc. The 157 observations are distributed among the four phases as follows (see Weihs et al. (1999)):

Table 1: **Economic variables**. The abbreviation ‘gr’ stands for growth rates with respect to last year’s corresponding quarter.

Abbreviation	**Economic variable**
Y	Real gross national product (GNP) (gr)
C	Real private consumption (gr)
GD	Government deficit
L	Wage and salary earners (gr)
X	Net exports
M1	Money supply M1
IE	Real investment in equipment (gr)
IC	Real investment in construction (gr)
LC	Unit labor cost (gr)
PY	Real gross national product price deflator (gr)
PC	Consumer price index (gr)
RS	Nominal short term interest rate
RL	Real long term interest rate

- Upswing: 59 observations
- Upper turning point: 24 observations
- Downswing: 47 observations
- Lower turning point: 27 observations

3 Univariate Outliers

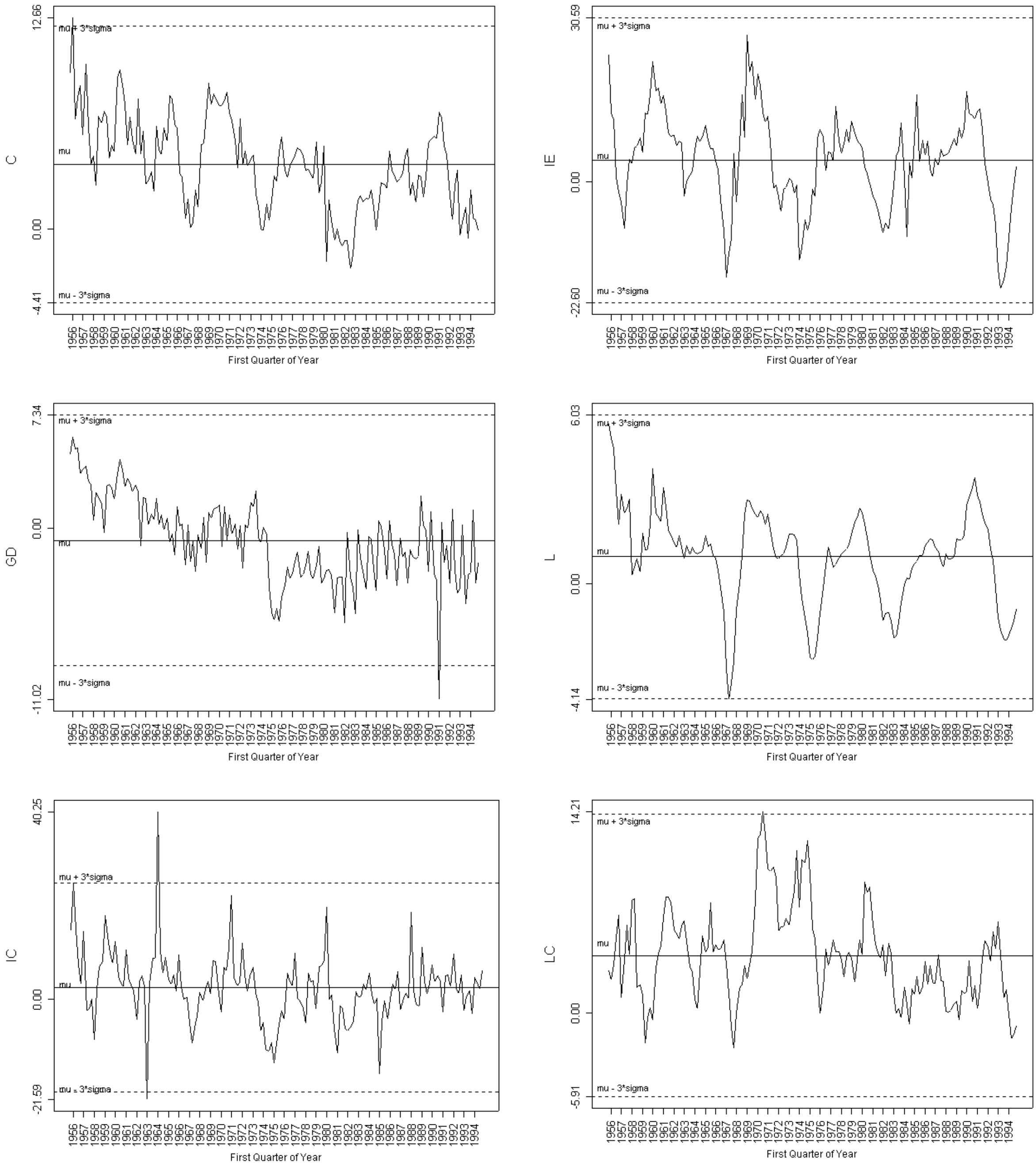
First of all we examine the data separately for each of the thirteen business factors. Thus, we treat the data as independent univariate time series in order to determine univariate outliers for each of those thirteen time series separately. A simple and common way of doing this is to plot the univariate time series and to mark mean μ and upper and lower bounds $\mu \pm 3\sigma$. Then, observations with values outside the bounds shall be indicated as possible outliers. Note that this procedure implicitly assumes that the times series is a realization of a normally distributed White Noise process $\varepsilon_t \sim N(\mu,\sigma^2)$, which is at least a questionable assumption in our case. The 3σ rule declares 8 observations as possible outliers, those are listed in Table 2.

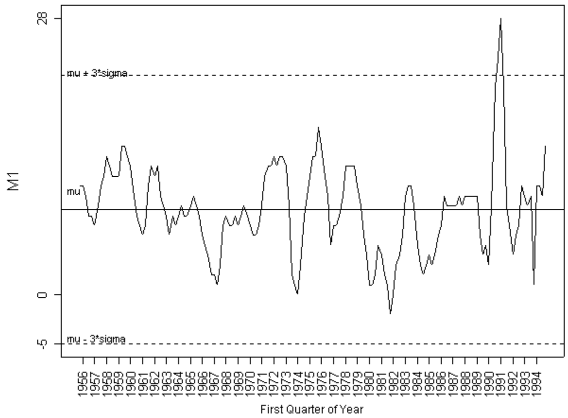
Table 2: **Possible univariate outliers determined by 3σ rule.** The highlighted observations are also recognized as potential outliers by the 4σ rule $\mu \pm 4\sigma$.

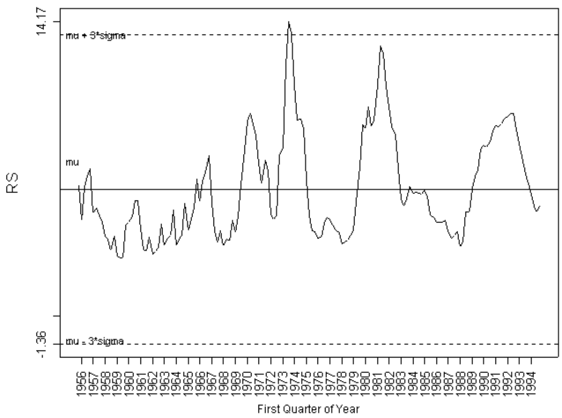
Economic variable	**Time point**	**μ - 3σ**	**μ + 3σ**	**Value**
C	1956/1 (observation 2)	-4.41	12.13	12.66
GD	1991/1 (observation 142)	-8.84	7.34	-11.02
IC	1963/1 (observation 30)	-19.88	25.01	-21.59
IC	**1964/1 (observation 34)**	**-19.88**	**25.01**	**40.25**
L	1967/2 (observation 47)	-4.11	6.03	-4.14
LC	1970/3 (observation 60)	-5.91	14.00	14.21
M1	**1991/1 (observation 142)**	**-5.00**	**22.23**	**28**
RS	1973/3 (observation 72)	-1.36	13.54	14.17

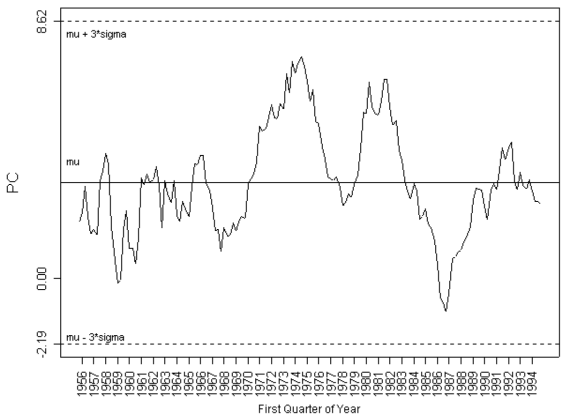
Observation 142 (1991/1) is marked as an outlier in both, variable GD (government deficit) and variable M1 (money supply M1). It is also remarkable, that 5 out of 8 potential outliers belong to a first quarter of a year, but none belong to a fourth quarter. Figure 1 shows the univariate time series plots with marked upper and lower bounds.

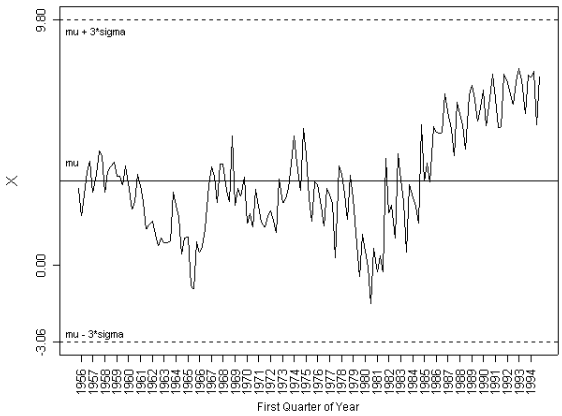
Figure 1: **Univariate time series for each of the 13 economic variables.** Observations, which exceed the bounds $\mu \pm 3\sigma$, are potential outliers according to the 3σ rule.

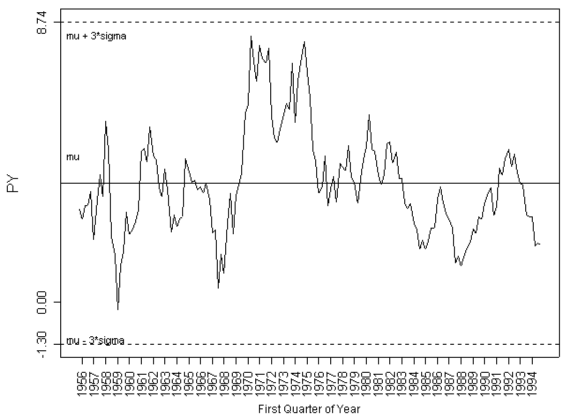


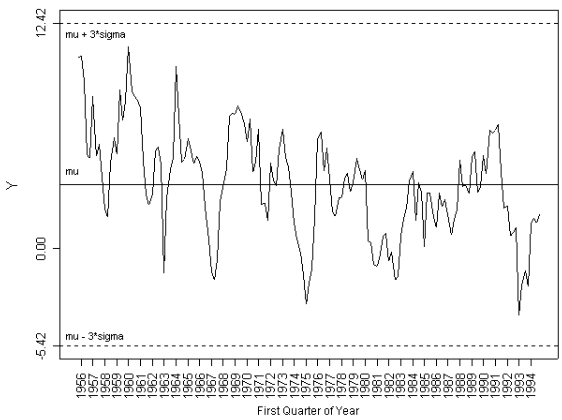

M1
28
0
-5
mu + 3*sigma
mu
mu - 3*sigma
1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994
First Quarter of Year

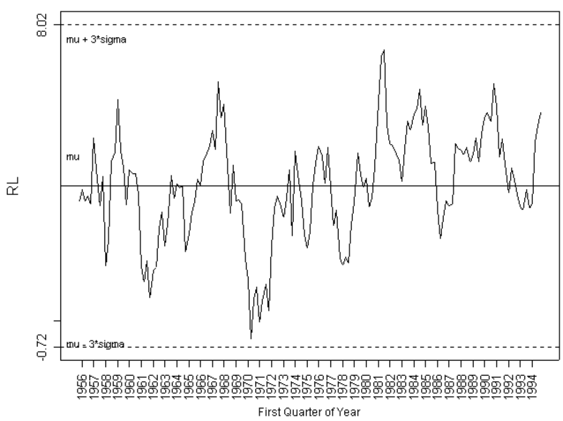

RS
14.17
-1.36
mu + 3*sigma
mu
mu - 3*sigma
1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994
First Quarter of Year


PC
8.62
0.00
-2.19
mu + 3*sigma
mu
mu - 3*sigma
1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994
First Quarter of Year


X
9.80
0.00
-3.06
mu + 3*sigma
mu
mu - 3*sigma
1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994
First Quarter of Year


PY
8.74
0.00
-1.30
mu + 3*sigma
mu
mu - 3*sigma
1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994
First Quarter of Year


Y
12.42
0.00
-5.42
mu + 3*sigma
mu
mu - 3*sigma
1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994
First Quarter of Year


RL
8.02
-0.72
mu + 3*sigma
mu
mu - 3*sigma
1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994
First Quarter of Year

4 Multivariate Outliers

Splitting a multivariate time series into univariate time series and treating those separately means, that the aspect of dependency between factors is ignored. For example, an extreme value in one component can be caused by an outlier in another component. This extreme value would not be an outlier on its own, but a univariate procedure would eventually determine it to be one. Besides, there can be outliers which affect all components, but do not cause very extreme values, so these outliers would not be detected by a univariate analysis.

One possibility to detect multivariate outliers is to fit a simple AR(1) process to the multivariate time series for all observations but one. If we leave out each of the 157 observations once (*leave-one-out* method), then the change in goodness of fit can be interpreted as a measure of "how well each of the observations fits into all the others". We measure the goodness of fit with the Akaike Information Criterion (AIC), where
AIC = -2loglikelihood + 2npar (npar = 'number of parameters' is equal for all models in this case, so just using the loglikelihood would give the same results). The smaller the AIC the better the model fits. If AIC_{-i} corresponding to observation i is unusually small, then leaving out observation i lead to a great improvement of model fit. This indicates that this observation "does not fit well to all the others". It might be a multivariate outlier.

Figure 2 illustrates the results graphically. It shows the ratios between the values AIC_{all} and AIC_{-i}, or rather the natural logarithms of those ratios:
$\log(AIC_{all}/AIC_{-i}) = \log(AIC_{all}) - \log(AIC_{-i})$.

In general, the values AIC_{-i} calculated from fitting only 156 out of 157 observations are smaller than the overall AIC_{all} value, therefore the model fit usually becomes better by leaving out an observation. Though there are no observations obvious with an extremely large log ratio, five observations have large values which can be clearly distinguished from the others. We mark those five observations as possible multivariate outliers, which have to be examined further. Observations that have been classified as potential univariate outliers, are also marked in Figure 2 in order to examine eventual overlaps. None of those univariate outliers have remarkable large log ratios of AIC values. Also, none of the possible multivariate outliers stand out as univariate extreme values. So, there is no overlap between multivariate and univariate outliers.

The five possible multivariate outliers are listed in Table 3. Again, a remarkably large proportion of values (three out of five) belongs to observations from first quarters of years.

Figure 2: **Multivariate outliers.** The 5 potential multivariate outliers are marked as filled triangles; Observations which have been classified as univariate outliers previously are marked as filled squares.

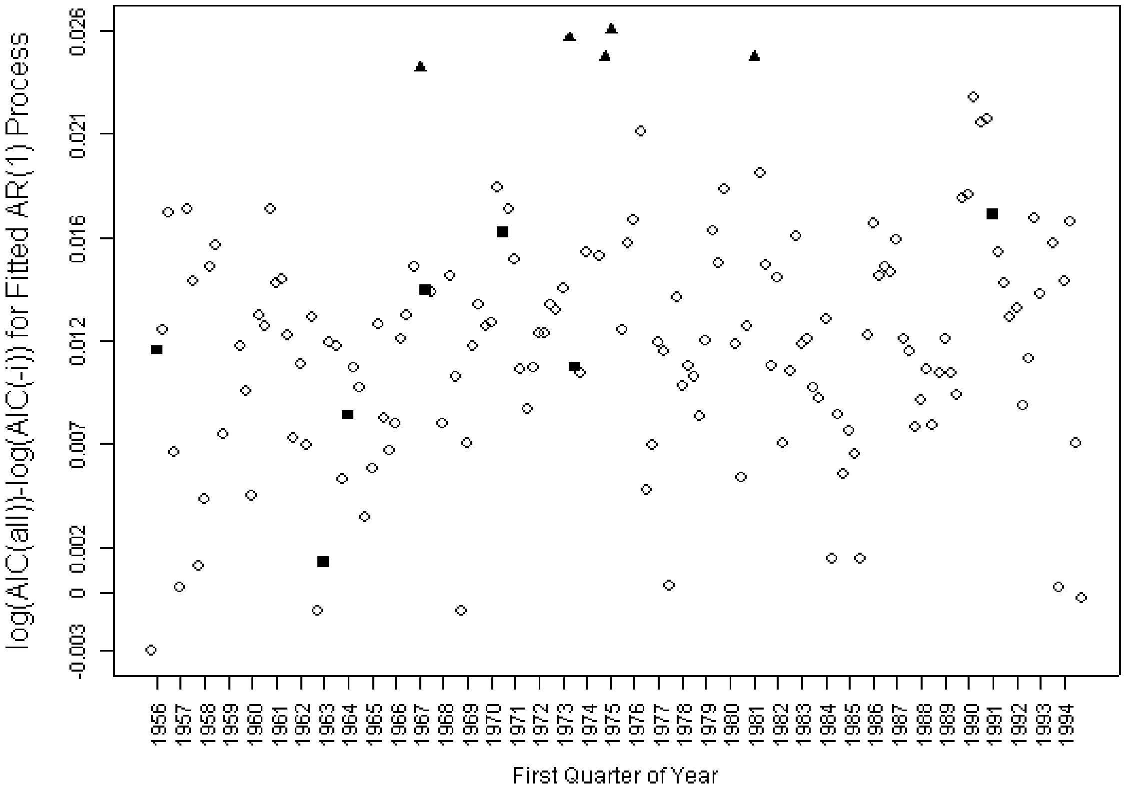


Table 3: **Possible multivariate outliers determined by fitting an AR(1) process.**

Observation	**Time point**	**AIC$_{-i}$**	**log(AIC$_{all}$) – log(AIC$_{-i}$)**
46	1967/1	1545.750	0.02435
71	1973/2	1543.678	0.02573
77	1974/4	1544.996	0.02484
78	1975/1	1543.083	0.02608
102	1981/1	1545.030	0.02482

Note: AIC$_{all}$ = 1583.857

5 Influential Observations

We can interpret the data as 157 observation vectors of length 13. These vectors can be classified into one of the four business cycle phases *upswing (1)*, *upper turning point (2)*, *downswing (3)* and *lower turning point (4)* (see section “The Data Set”). The Linear Discriminant Analysis (LDA) automatically classifies observations into one of some known classes using a classification rule. In our case those known classes are given by the business cycle phases.

We can use the Linear Discriminant Analysis to determine influential observations by performing cross validation using the *leave-one-out* method. The coefficients of the linear discriminant functions, that are obtained, if all but observation i are part of the training data set, shall be named coeff_{-i}. If those coefficients differ much from all the other coefficients coeff_{-j} ($j \neq i$), then observation i might be an influential observation. The discriminant analysis with all 157 observations in the training data set gives the following discriminant function coefficients.

Table 4: **Coefficients of discriminant functions for LDA with all observations.**

Economic variable	Standardized canonical coefficients of discriminant functions		
	Function		
	1	**2**	**3**
IE	-0.413	0.194	0.365
C	0.785	-0.027	0.935
Y	-0.375	-0.689	-0.582
PC	0.459	-0.459	-0.075
PY	-1.467	1.760	1.454
IC	0.032	0.296	0.123
LC	0.458	-0.090	-0.872
L	-0.508	-0.842	-0.137
M1	0.476	0.190	0.415
RL	-0.961	1.317	0.976
RS	1.535	-0.844	0.040
GD	0.551	0.594	0.254
X	-0.410	0.010	-0.174

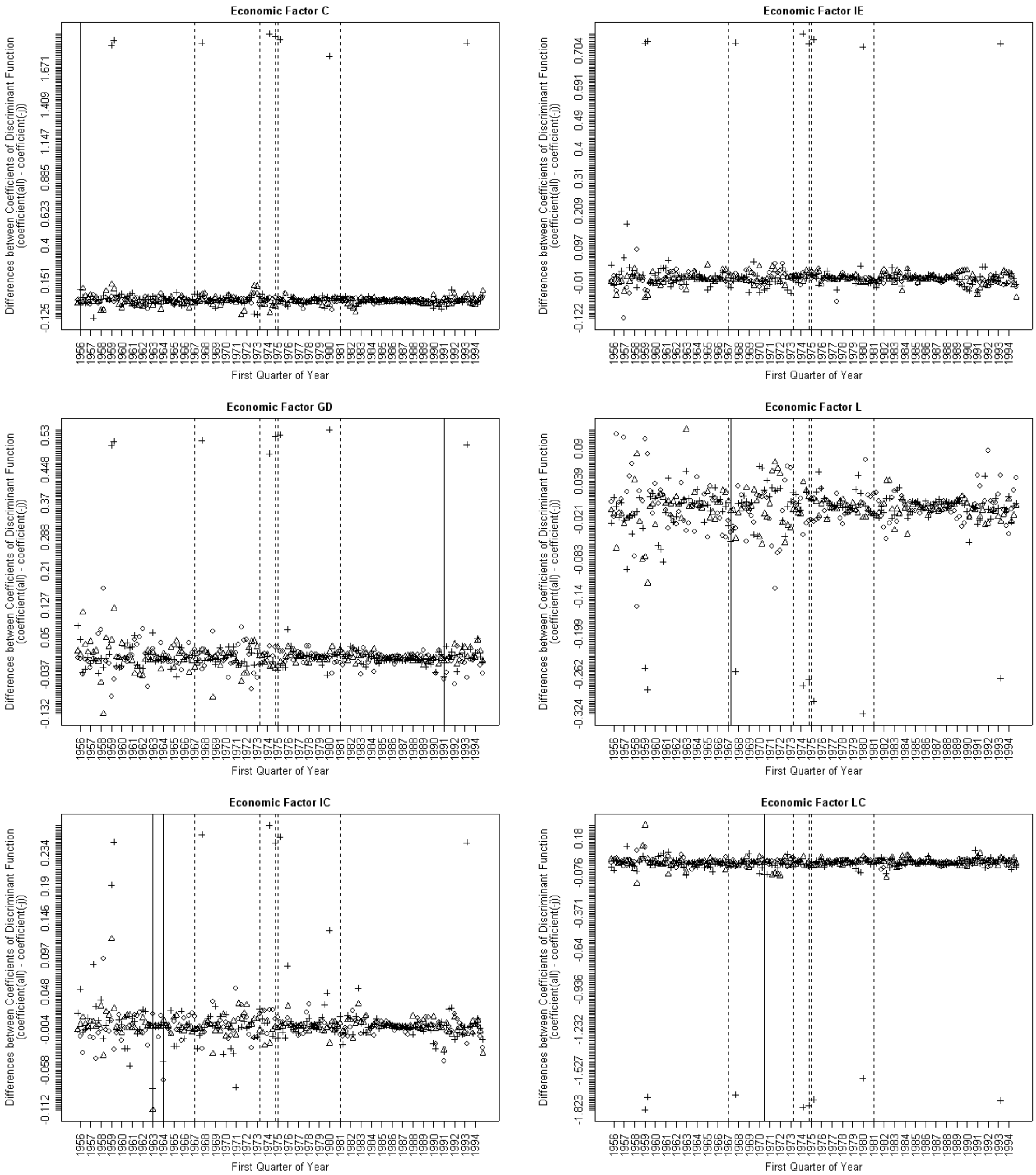
If S represents the pooled covariance matrix within classes and B is the between-classes covariance matrix, then the eigenvalues of matrix $S^{-1}B$ correspond to the first three discriminant functions. The proportion of trace of matrix $S^{-1}B$, which is explained by an eigenvalue, correlates to the portion of variance, which is explained by the corresponding discriminant function. The eigenvalues and variance proportions can be found in Table 5.

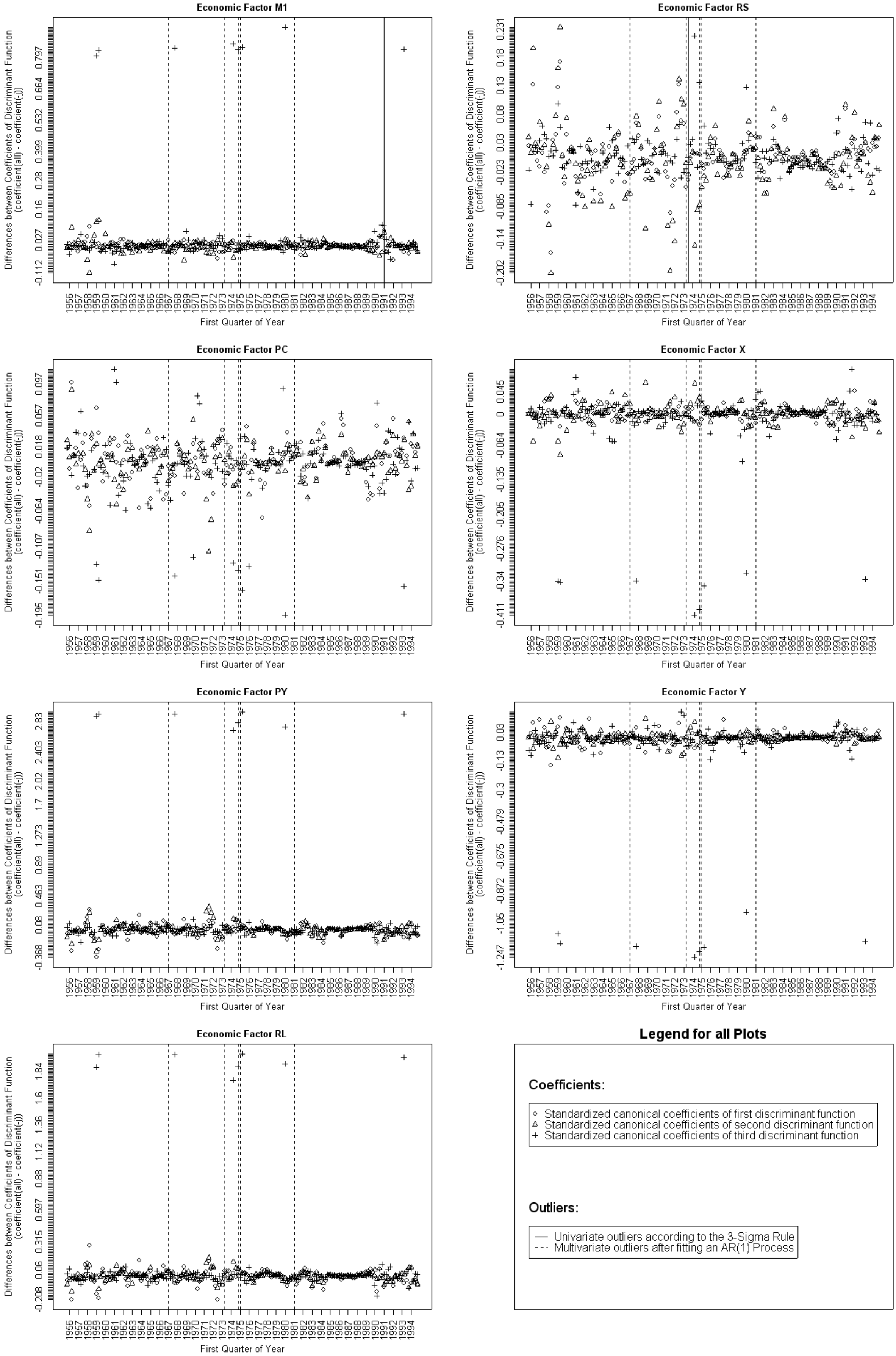
Table 5: **Eigenvalues and corresponding proportions of variance.**

Function	Eigenvalue	Proportion of variance	Cumulated proportion of variance
1	1.512	0.527	0.527
2	1.091	0.380	0.907
3	0.266	0.093	1.000

An intuitive way of illustrating the differences coeff_{all} - coeff_{-i} graphically would be to create three graphics, one for each discriminant coefficient function, and to plot - at each time point - the values of each of the 13 economic variables. But since these plots would contain too many points, it would become hard to detect anything. So we create 13 graphics, one for each economic factor, with differences in coefficients coeff_{all} - coeff_{-i} for all three discriminant functions plotted at each time point. Additionally, vertical lines mark those observations, which have been determined to be either univariate or multivariate potential outliers. The results are shown in Figure 3.

Figure 3: **Influential observations determined by LDA.** The dotted vertical lines indicate the five observations that have been classified as possible multivariate outliers; the solid lines indicate potential univariate outliers.





It is striking that for nearly all economic variables – except for PC (consumer price index) and RS (nominal short term interest rate) – exactly the same eight observations differ extremely from all other observations in the third discriminant function coefficients. We indicate these eight observations as influential. The values of the third discriminant function $coeff_{-i}$ of those eight extreme observations are listed in Table 6. Table 7 contains the corresponding values of the differences $coeff_{all}$ - $coeff_{-i}$. Contrary to the univariate and multivariate outliers, observations from the first quarter of a year are not over-represented here.

Table 6: **Influential observations determined by LDA** and extreme values of the standardized canonical coefficients of the 3rd discriminant functions. The business cycle phases are the actual phases, determined by the experts' classification (Heilemann and Münch (1996)).

Economic variable	**Observation number**							
	14	**15**	**49**	**75**	**77**	**79**	**98**	**151**
	Time point (year/quarter)							
	1959/1	**1959/2**	**1966/4**	**1974/2**	**1974/4**	**1975/2**	**1980/1**	**1993/2**
Business cycle phase	4	4	4	4	4	4	2	3
IE	-0.361	-0.366	-0.361	-0.390	-0.358	-0.371	-0.348	-0.359
C	-0.903	-0.934	-0.921	-0.985	-0.968	-0.943	-0.825	-0.921
Y	0.531	0.588	0.605	0.665	0.632	0.608	0.410	0.576
PC	-	-	-	-	-	-	-	-
PY	-1.440	-1.475	-1.475	-1.248	-1.358	-1.504	-1.302	-1.473
IC	-0.068	-0.126	-0.136	-0.149	-0.125	-0.133	-0.007	-0.126
LC	0.951	0.859	0.843	0.929	0.920	0.876	0.718	0.885
L	0.116	0.149	0.121	0.144	0.132	0.168	0.187	0.131
M1	-0.387	-0.413	-0.422	-0.439	-0.416	-0.424	-0.508	-0.415
RL	-0.896	-1.008	-1.011	-0.781	-0.899	-1.019	-0.931	-0.983
RS	-	-	-	-	-	-	-	-
GD	-0.251	-0.261	-0.262	-0.232	-0.274	-0.276	-0.289	-0.253
X	0.169	0.170	0.166	0.237	0.224	0.176	0.150	0.164

Table 7: **Eight most extreme differences $coeff_{all}$ - $coeff_{-i}$** between standardized canonical coefficients of 3rd discriminant functions for all economic factors.

Economic variable	**Observation number**							
	14	**15**	**49**	**75**	**77**	**79**	**98**	**151**
	Time point (year/quarter)							
	1959/1	**1959/2**	**1966/4**	**1974/2**	**1974/4**	**1975/2**	**1980/1**	**1993/2**
IE	0.726	0.731	0.725	0.754	0.722	0.736	0.713	0.724
C	1.839	1.870	1.856	1.921	1.903	1.879	1.760	1.856
Y	-1.113	-1.169	-1.186	-1.247	-1.214	-1.190	-0.992	-1.158
PC	-	-	-	-	-	-	-	-
PY	2.893	2.928	2.928	2.701	2.811	2.958	2.756	2.926
IC	0.191	0.249	0.259	0.271	0.248	0.256	0.130	0.248
LC	-1.823	-1.731	-1.715	-1.801	-1.792	-1.748	-1.590	-1.757
L	-0.253	-0.286	-0.258	-0.281	-0.269	-0.305	-0.324	-0.268
M1	0.802	0.828	0.837	0.854	0.831	0.840	0.923	0.830
RL	1.873	1.985	1.987	1.757	1.876	1.996	1.907	1.960
RS	-	-	-	-	-	-	-	-
GD	0.505	0.516	0.517	0.486	0.528	0.530	0.543	0.507
X	-0.343	-0.344	-0.340	-0.411	-0.398	-0.350	-0.324	-0.337

6 Discussion and Comparison

We applied several methods in order to get an impression about extreme or influential observations in the German Business Cycle data set. Extreme or influential observations in a multivariate time series can have different characters. They can be either univariate or multivariate outliers, or influential points. We used a simple 3σ rule to determine possible univariate outliers. For getting an impression on potential multivariate outliers, we fitted an AR(1) process to the data, and in order to find influential points we performed a Linear Discrimant Analysis.

It is remarkable, that there have been nearly no overlaps between the different types of extreme observations. Specifically, there are none between the sets of univariate and multivariate outliers. Only observation 77 (1974/4) is classified as a multivariate outlier, and an influential observation. But there do exist clusters, where several extreme values accumulate. The biggest cluster lies between years 1973 and 1975, where three out of five potential multivariate outliers, one univariate outlier, and three out of eight influential observations are located. This corresponds to the period of the first Mideast oil crisis, which started in the end of 1973, when the Arab member countries of OPEC drastically restricted their crude oil production, which had a large impact on Germany's energy policy. While most of the multivariate outliers lie in this time period, many univariate outliers can be found in the beginning of the observation period - in the 1960´s. One reason for those may be the economic upswing in West Germany during the 1960´s. An interesting economic variable in this context is investment in construction (IC), which has two univariate outliers in 1963 and 1964. In 1963/1 there was a sharp decline in investment, but in 1964/1 the investment in construction was increased dramatically – maybe as a counter maneuver. There is one single univariate outlier in the first quarter of 1991, meaning neither belonging to the 1970's cluster nor to the extreme values from the beginning of the observation period. This outlier occurred in factor M1 (money supply M1). It can be easily explained by the German reunification, which took place in 1990, and the simultaneous monetary reform. A remarkably large proportion of all univariate (5 out of 8) and multivariate (3 out of 5) outliers have been obtained at a first quarter of a year.

The influential observations accumulate in both mentioned time periods – in the seventies (1974 and 1975), and in the beginning of the observation period (1959). Interestingly, six out of eight detected influential points actually belong to the business cycle phase *lower turning point* (phase 4), whereas only a relatively small proportion of all observations are classified into this category (27 out of 157).

7 Conclusion

We used simple methods to determine univariate and multivariate outliers, and influential observations in a multivariate time series. The outliers, that have been detected by those methods, can well be explained by economic and historic facts. It is remarkable, that there are nearly no overlaps between the different types of extreme values. For example, if an observation has extreme values in one of the variables, then this does not necessarily have an impact on its overall performance in all variables. It can still fit well to all other observations in the multivariate context. But, on the other hand, the extreme values cluster in specific time periods, especially during the first Mideast oil crisis 1973 to 1975. Therefore, extreme circumstances cause extreme and influential values in all kinds, whether as univariate or multivariate outliers, or as influential observations. An interesting fact is, that most influential points belong to the business cycle phase *lower turning point.*

References

Heilemann, U. and Münch, H.J. (1996). West German Business Cycles 1963-1994: A Multivariate Discriminant Analysis. CIRET-Conference in Singapore. CIRET-Studien 50. Munich, 220-250.

Tsay, R.S., Peña, D., Pankratz, A.E. (2000). Outliers in multivariate time series. Biometrika. **87**. 4. 789-804

Weihs, C., Röhl, M.C., Theis,W. (1999). Multivariate Classification of business phases. Technical Report 26/1999. SFB 475 at Fachbereich Statistik, Universität Dortmund.

Acknowledgement

This work has been supported by the Deutsche Forschungsgemeinschaft (DFG), SFB 475.

We thank Dr. D. Enache for valuable discussions about the economic interpretation.