Finnish Economic Papers – Volume 14 – Number 1 – Spring 2001

OUTLIERS IN ELEVEN FINNISH MACROECONOMIC TIME SERIES*

JUSSI TOLVI

Department of Economics, 20014 University of Turku, Finland

Outliers are detected from eleven monthly Finnish macroeconomic time series (from 1922 to 1996), and underlying reasons for the detected outliers suggested. Outliers are found in all of the series. The detected outliers seem to cluster together, both within and across series. Some of the detected outliers are also highly influential with respect to some nonlinearity tests, and the results can change in either direction (from significant nonlinearity to non-significant, or vice versa) after the outliers are taken into account. Mostly, however, the evidence for nonlinearity is reduced after the outliers are taken into account. (JEL: C5, E3.)

1. *Introduction*

Outliers have theoretically been shown to have adverse effects in many situations in time series analysis. Examples of these include linear ARMA models (Tsay 1986a, Ledolter 1989, Chen and Liu 1993), ARCH tests and models (van Dijk, Franses and Lucas 1999a, Franses and van Dijk 1999, Sakata and White 1998, Tolvi 2000), tests for unit roots and cointegration (Franses and Haldrup 1994, Hoek, Lucas and van Dijk 1995) and nonlinearity tests and nonlinear models (smooth transition autoregression in van Dijk, Franses and Lucas 1999b and bilinearity in Chen 1997, Gabr 1998 and Tolvi 2000). In all these the presence of outliers may distort the results of testing and estimation.

Of course, such results are based on theoretical arguments, and for them to have practical relevance empirical economic data should contain outliers (that is, observations that are in the theoretical literature defined as outliers with respect to a certain model). Examining this question brings up some problems, since the empirical data and the theoretical models do not necessarily agree. Nevertheless, some work has been done, assuming that the data can be reasonably approximated with a linear autoregressive model. Balke and Fomby (1994) examine quarterly (and also a few monthly) U.S. macroeconomic time series, and find a large number of significant outliers. Here I will do a similar search for some Finnish macroeconomic series, but with some improvements in the outlier detection method, and also for longer, monthly time series. The methodological improvements are the inclusion of a new type of outlier, which provides a more varied classification of aberrant events, and the use of an improved outlier detection algorithm.

* *I would like to thank Charles Bos, Dick van Dijk, Matti Virén, seminar participants at the Econometric Institute of Erasmus University Rotterdam and two anonymous referees for comments, and Kari Takala and Matti Virén for providing me with the data.*

The effects of the detected outliers are then examined by some diagnostic measures and nonlinearity tests. It is often found that nonlinearity tests reject the hypothesis of linearity for economic time series. According to the theoretical literature mentioned earlier, these rejections might not be due to nonlinearity alone, but also to outliers that have not been taken into account. To determine which of these is the more probable cause for the rejection of linearity, I will in this paper compute the tests both for the original series and for series where the effects of the outliers have been removed. If nonlinearity is detected from these outlier corrected series as well, there is more reason to believe that there really is nonlinearity in the data. If, however, the tests do not reject linearity for the outlier corrected series, it can be concluded that the rejection of linearity was perhaps due to the outliers. Nevertheless, there are also some difficulties with this reasoning, since it is possible that the outlier detection method will detect nonlinearities as well. Therefore, to some degree at least, the decision has to be based on the analyst's judgement. The number of detected outliers is important, since outliers occur, by definition, only rarely. Therefore, if taking only a few outliers into account changes the results considerably, it seems appropriate to conclude that these outliers are the probable cause of the detected nonlinearity. If, on the other hand, a larger proportion of the data are classified as outliers, a more convincing explanation may be nonlinearity of some kind. It is perhaps worth noting, that some nonlinearity tests may be highly sensitive to outliers. Even a single outlier in several hundred observations may distort the tests' sizes considerably. Tolvi (2000) shows this for ARCH and bilinearity tests.

The data to be used in this paper will first be presented in section 2, followed by definitions for outliers and a brief description of the outlier detection method. Some earlier empirical work is also reviewed. The results of the outlier detection are given in section 3, along with a discussion of the outliers' possible causes. Section 4 presents the diagnostics and the nonlinearity tests that are used for the data, and the results for both the original data and for the outlier corrected data. Section 5 concludes with a brief discussion on the implications of the results.

2. Data and methods

2.1 Data

The data series are the same as were used in Takala and Virén (1997). The eleven variables are industrial production (IP), number of bankruptcies (BA), real exchange rate (FX), terms of trade (TT), consumer price index (CPI), wholesale price index (WPI), banks' total credit supply (LEND), narrow money supply (M1), broad money supply (M2), UNITAS stock exchange index (SX) and turnover in Helsinki stock exchange (ST). The IP, BA, M1 and M2 series are seasonally adjusted. The data is monthly, and covers the period from January 1922 to September 1996, $T = 897$. Note that the original data set of Takala and Virén also included an interest rate variable which is omitted here, since the interest rates in Finland were administratively fixed during most of this time period. For more details on the data, see Takala and Virén (1997) and the references given there.

Logarithmic differences of the series are examined, apart from the exchange rate and terms of trade series, for which differences are used. These transformations were also used by Takala and Virén (1997). The analysis was carried out in two subsamples, the first from 1922:2 to 1963:9, and the second from 1955:2 to 1996:9. Both subperiods have 500 observations. The two subperiods will be distinguished by numbers attached to the series name, so that for example IP1 and IP2 refer to the industrial production series in the first and second subperiods, respectively. There are two reasons for this split of the data. First, it is likely that the series have undergone structural changes during the seven decades of observations. And if the linear model fits badly, the outlier detection method may encounter problems in detecting and identifying the outliers correctly. By splitting the data the parameters of the linear model, and also the variances of the error terms are allowed to differ for the two subperiods, thus diminishing the effects of possible structural changes.

Second, since the two periods overlap, it is interesting to examine whether the results for the two outlier detections agree in the overlapping period (i.e., it will be seen whether the same outliers are detected from both subperiods).

An AR(4) model is used as the linear model in the following analysis for all series in both periods (it was also used by Takala and Virén). This may not be an optimal choice in every case, but it seems to remove most of the linear dependence in the series.

2.2 Methods

Definitions of the outliers examined here will be given next. For details, see Tsay (1986a and 1988) and Chen and Liu (1993). The definitions are given only for autoregressive (AR) models, since they are the only ones used here. For general formulations for ARMA models, see the references given above.

The observed series x_t is composed of the underlying series y_t and three types of outliers, such that

(1) $x_t = y_t + \omega_{AO,j}\, I[t = j] + \lambda^c \omega_{TC,k}\, I[t \geq k] + \omega_{LS,l}\, I[t \geq l]$.

Here $\omega_{AO,j}$ are the additive outlier (AO) magnitudes, and $I[t = j]$ is an indicator variable which is equal to one when $t = j$ and zero otherwise, and the outlier therefore affects only one observation. $j = \{j_1, j_2, \ldots, j_A\}$ are the timings of the additive outliers. The next outlier type is the temporary change (TC) outlier (of magnitudes $\omega_{TC,k}$), for which $I[t \geq k]$ is equal to one for all $t = k + c$, $c = 0, 1, 2, \ldots$, or $t \geq k$, and zero otherwise. Since it is also required that $0 < \lambda < 1$, the effect of a TC dies out exponentially. Usually a value of 0.7 is chosen for λ, which is used here as well. $k = \{k_1, k_2, \ldots, k_C\}$ are the timings of the TCs. The third outlier type is the level shift (LS) outlier (of magnitudes $\omega_{LS,l}$), for which $I[t \geq l]$ is also equal to one for all $t \geq l$ and zero otherwise, in other words the effect of the outlier is permanent. Again, $l = \{l_1, l_2, \ldots, l_S\}$ are the timings of the level shifts. Strictly speaking, a level shift is not an outlier, but rather a structural change. It is nevertheless often included as an outlier in the literature, and will be initially considered here as well (see the discussion in section 3).

The underlying series y_t is an AR series with the fourth type of outliers, such that

(2) $\phi(L)y_t = \varepsilon_t + \omega_{IO,m} I[t = m]$,

where $\phi(L)$ is a lag polynomial of degree p (here $p = 4$), ε_t is an i.i.d. normally distributed disturbance process, $\omega_{IO,m}$ are the innovative (or innovational) outlier (IO) magnitudes and $I[t = m]$ an indicator variable similar to the one of an AO. Because of the AR part in this equation, however, the effect of an IO is carried on to the future observations as well. $m = \{m_1, m_2, \ldots, m_I\}$ are the timings of the IOs.

There are several methods for the detection of the outliers defined above. An often used procedure is that of Tsay (1988). This method was also used by Balke and Fomby (1994), although with some modifications (see later). Here I will use an improved algorithm by Chen and Liu (1993), which is readily available, with slight modifications, in the computer program TRAMO (Gómez and Maravall 1994, 1995). The detection is based on adding appropriate intervention variables (corresponding to the four outlier types given above) for every observation in turn, computing simplified likelihood ratio statistics (compared to a model without the interventions) and finding the significant ones. In the first stage of the detection procedure, an ARMA model is estimated assuming there are no outliers in the data. The estimated residuals are then examined, and likelihood ratio test statistics for the four types of outliers calculated for each observation. The largest absolute value of these test statistics is compared to a prespecified critical value, and if the test statistic is larger, an outlier is found. The effect of this outlier is removed from the data, and this process repeated until no more outliers can be found. The algorithm returns to the beginning and the ARMA model is re-estimated, outlier tests are repeated until no outliers can be found, and this outer loop is repeated until no outliers are found in the inner loop.

In the second stage, the effects of the detected outliers are estimated and standardized. These standardized values are then compared to the same critical value as in stage one, and if the estimated outlier effect is not significant at this stage, the outlier in question is dropped from consideration. This step is repeated starting from the outlier with the smallest test statistic value, and stopped when all values are larger than the critical value. The series is then adjusted using the remaining outlier effects, and the model is re-estimated. If the relative change in the residual standard error is larger than a predetermined constant, the outlier effects are re-estimated and insignificant outliers deleted iteratively until the change is negligible. The series is now filtered with the remaining outlier effects, and the iteration returns to the beginning of the first stage. The ARMA model is again estimated, and the outlier search and deletion stages repeated. The algorithm stops when after re-estimating the ARMA parameters no outliers are either found or deleted in the following stages. The outlier corrected data is then achieved simply by removing from the original data the effects of the detected outliers.

In TRAMO different critical values for the outlier detection can be used to vary the sensitivity of the search. Indeed, one of the drawbacks of these methods is that there is no formal theory behind the critical values. They are selected based on informal arguments and some simulation results. Here a critical value of 4.0 is used, which is considered a low sensitivity value for this sample size. This choice and its consequences will be discussed later in some more detail. Unless otherwise noted, the default options of TRAMO are used in this paper.

It should perhaps also be kept in mind that temporal aggregation will usually reduce the nonlinearity of observed time series, and the same is true of outliers as well (see, e.g. Granger and Teräsvirta 1993). Less outliers should therefore be detected in annual and quarterly macroeconomic data than in monthly data. This was also found by Balke and Fomby (1994), who note, however, that the relative number of outlier observations (out of all observations) is not much greater in monthly data than in quarterly data.

2.3 Earlier results

Balke and Fomby (1991 and 1994) have studied some U.S. macroeconomic series in the light of the recent literature on outliers (note that they detected outlier types AO, IO and LS only). In Balke and Fomby (1991) they considered GNP and GNP deflator data, and compared unit root, shifting trend and outlier models for these series. In their conclusions they state that "(t)here is strong evidence that just a few important events or shocks [i.e. outliers] account for most of the persistence in the GNP deflator", but they did not find similar results for real GNP. In Balke and Fomby (1994) they analyzed 15 quarterly and three monthly macroeconomic series (1947–1992). Almost all of these series have significant skewness and (excess) kurtosis, GARCH effects and nonlinearity in their AR residuals. Several significant outliers are detected in most series. In the quarterly series half of the detected outliers are of the AO type, somewhat more than a quarter are IOs, and the rest LSs. In the monthly series more IOs and less LSs are detected than in the quarterly series. When the series (both quarterly and monthly) are corrected for these outliers, evidence of nonlinearities and non-normality of their AR residuals is considerably decreased, and even disappears completely for many series.

From an economics perspective, Balke and Fomby (1994) find some interesting general patterns in the detected outliers. First, many of the outliers occur at the turning points of the business cycle or during recessions. Second, the detected outliers are often clustered together. Often several outliers occur close to one another in a single series. This phenomenon, which is typical especially in price series, may be due to changes in volatility and ARCH type behavior. Also, in many instances an outlier is detected at the same time in several series, which suggests that the outliers may have common causes. Third, there is a notable difference in real (production, consumption, unemployment etc.) and nominal (price, wage etc.) series. The amount and influence of detected outliers is considerably smaller in the real series, whereas the nominal series have more outliers, which also change the properties of the series more substantially. In addition, Balke and Fomby tried to identify the underlying causes of the detected outliers. This is of course a somewhat speculative enterprise, although in many cases the identified events are quite reasonable (e.g. the effects of a steel strike on industrial production).

Note finally, that the outlier detection procedure used by Balke and Fomby is based on the algorithm of Tsay (1988) with some modifications, and the critical value used was 3.0, which is considered a medium sensitivity value. The Tsay procedure does not consider dropping detected outliers if they become insignificant at a later stage, and thus it is possible that too many outliers are detected. Balke and Fomby protected their results against this possibility by re-estimating the models with outlier interventions, examining the detected outliers' *t* values and dropping any insignificant ones from the final model. This is approximately equivalent to stage two in the method used here.

In an interesting study, Fischer and Planas (1998) examined a set of over 13 000 monthly series of industrial production, turnover, new orders, imports and exports, mostly from countries in the European Union. The majority of the series have less than or about 100 observations. Among other results they found an average of 1.25 outliers per series; more in the industrial production series, less in the foreign trade series. Additive outliers were found in about 40% of the series, temporary changes and level shifts in a little over 20% (IOs were not searched for). Only less than 40% of the series did not have any outliers. Of course these numbers depend on the critical value used in the outlier detection procedure. Fischer and Planas used rather large (low sensitivity) critical values, and with smaller ones the amount of detected outliers would naturally have been greater. Note also, that as in this paper, Fischer and Planas used the Chen and Liu (1993) outlier detection procedure, as implemented in the program TRAMO.

Regarding the data set used here, Takala and Virén (1997) did an extensive analysis of several aspects of it. They found, first of all, that the data are clearly non-normal and nonlinear. Deterministic chaos, on the other hand, is not implicated, whereas some evidence of long memory is found. On the possible form of the detected nonlinearity, Takala and Virén conclude that ARCH is not a likely candidate. Instead, they favor asymmetric models that are able to capture the adjustment processes involved in the business cycle. This conclusion is enforced by the finding that many series behave differently in recessions and expansions.

3. *Outlier detection*

Outliers were first detected from the data, using the method described earlier. In the first round, when the AR(4) model was used, no level shifts were detected in any of the series. In addition to a genuine lack of level shifts in the data, this may also be due to the fact that the outlier detection procedures seem to have low power in detecting level shifts, and a tendency to misidentify them as innovational outliers. This possibility is discussed by Balke (1993). He therefore suggests that the outlier detection procedure should also be run using an ARMA(0,0) model as the initial model, which increases the power of level shift detection. When this is done, some level shifts are indeed detected in some series (namely in CPI2, WPI2, LEND1 and LEND2 and two in M21). According to Balke, the ARMA(0,0) detection procedure may also detect spurious level shifts, due to neglected autocorrelation, and therefore any detected level shifts should be tested before they are accepted in the final model. In the following these detected level shifts were therefore included in the AR(4) model as intervention variables, and the outlier detection repeated for this model. In these results, all of the earlier detected level shifts remained statistically significant, and were therefore included in the final models. It may be of some interest to note, that compared to the results obtained by excluding the level shifts, there were some changes in the detected outliers when the previously detected level shifts were included in the models. For three series (CPI2, LEND1 and M21), some additional outliers were detected, for LEND1 two outliers were not detected any more, and for LEND2 one outlier's identification changed from an AO to an IO. For WPI2 there were no changes in the detected outliers.

The final results of the outlier detection are given in tables 1 and 2. In Table 1, all detected

Table 1. Detected outliers by series.[1]

Series	Type	Date	Size	S.E.	Event
Industrial production	TC	1939:10	–0.15	0.030	Onset of war in Europe
	AO	1941:6	–0.24	0.058	War
	IO	1950:9	–0.32	0.065	Strikes
	IO	1950:11	0.28	0.065	End of strikes
	IO[3]	1955:7	–0.20	0.026	
	TC[3]	1956:4	0.075	0.014	End of strikes
	IO[3]	1956:7	–0.23	0.027	
	IO[3]	1958:9	–0.17	0.026	
	TC[3]	1958:12	0.073	0.013	
	IO[3]	1959:9	–0.11	0.026	
	IO	1967:7	0.11	0.026	
	IO	1968:7	0.10	0.026	
	AO	1970:1	–0.13	0.024	
	AO	1971:2	–0.25	0.024	
	AO	1971:4	0.27	0.024	
	AO	1977:3	–0.11	0.024	
Bankruptcies	IO	1939:12	–1.79	0.34	War
	IO	1941:7	–1.58	0.34	War
	IO	1943:9	–1.54	0.34	War
	IO	1945:11	–1.50	0.34	
	IO	1964:5	–0.89	0.20	
Consumer price index	AO	1923:8	0.047	0.011	
	AO	1942:8	0.063	0.011	Peak of wartime inflation
	TC	1945:6	0.16	0.0098	End of war
	AO	1945:10	0.13	0.011	Postwar inflation
	IO	1947:3	0.076	0.011	End of some food rationings
	TC	1947:11	0.070	0.0098	Wage increases
	IO	1948:2	–0.048	0.011	
	AO	1950:6	0.064	0.011	Rent and food price increases
	AO	1954:11	–0.062	0.011	Food price reductions and elimination of some sales taxes
	TC[3]	1956:1	0.021	0.0037	
	AO[3]	1956:11	0.034	0.0043	
	AO	1964:1	0.029	0.0043	
	AO	1967:1	–0.025	0.0043	
	IO	1967:12	–0.042	0.0044	
	IO	1968:1	0.069	0.0044	
	AO	1968:3	0.011	0.0043	
	AO	1972:1	–0.017	0.0043	
	AO	1973:7	0.017	0.0043	
	IO	1974:2	0.021	0.0044	
	TC	1975:1	0.018	0.0037	
	LS	1991:3	–0.0040	0.0012	
Wholesale price index	TC	1922:10	–0.032	0.0078	
	AO	1927:1	0.095	0.0086	Wood & paper price increases
	TC	1931:10	0.046	0.0090	Import price increases
	TC	1931:11	0.042	0.0088	Import price increases
	TC	1932:2	–0.039	0.0079	Reductions in import prices
	TC	1939:9	0.042	0.0078	Onset of war in Europe
	AO	1941:2	0.043	0.0086	War
	IO	1942:1	0.037	0.0091	War
	TC	1945:6	0.095	0.0080	End of war
	IO	1945:8	0.088	0.0092	Postwar inflation
	AO	1945:10	0.073	0.0086	Postwar inflation
	AO	1946:4	0.077	0.0086	Import price increases
	IO	1947:3	0.042	0.0091	End of some food rationings

continued

Table 1. Continued.

Series	Type	Date	Size	S.E.	Event
	AO	1947:8	0.051	0.0086	End of meat rationing
	IO	1947:11	0.080	0.0091	Wage increases
	AO	1951:1	0.11	0.0086	Industrial price increases
	AO	1954:11	–0.040	0.0086	Removal of sales taxes
	AO[3]	1956:11	0.019	0.0049	
	AO[2]	1957:10	0.038	0.0086	
	AO[3]	1957:10	0.037	0.0049	
	TC	1967:10	0.034	0.0046	
	IO	1973:7	0.027	0.0051	
	TC	1974:1	0.024	0.0046	
	TC	1976:7	0.023	0.0046	
	LS	1985:4	–0.0039	0.0012	
Bank lending	AO	1940:1	–0.061	0.0091	War
	AO	1940:7	0.041	0.0091	War
	AO	1941:1	0.052	0.0091	War
	AO	1941:7	0.042	0.0089	War
	AO	1941:9	0.044	0.0092	War
	IO	1942:3	0.051	0.0099	War
	TC	1942:7	–0.041	0.0089	War
	TC	1943:6	0.039	0.0089	War
	AO	1944:9	–0.067	0.0093	War
	AO	1944:10	0.046	0.0093	War, end of
	LS	1944:10	0.0098	0.0026	hostilities in Finland
	IO	1946:1	0.062	0.0099	
	IO[3]	1957:4	–0.022	0.0051	
	IO	1985:1	0.028	0.0051	Liberalization of financial markets
	AO	1988:12	0.036	0.0048	Tax changes
	AO	1991:11	0.021	0.0049	
	LS	1991:12	–0.014	0.0022	
	IO	1993:4	–0.023	0.0051	Plan to cut public spending
	IO	1995:1	0.025	0.0051	Finland joins the EU
M1	TC	1939:10	0.12	0.016	Onset of war
	IO	1973:4	0.18	0.024	
	IO	1973:12	0.11	0.024	
	IO	1975:12	0.11	0.024	
	IO	1978:5	0.11	0.024	
	AO	1989:2	0.085	0.022	
	AO	1990:1	0.21	0.022	
	AO	1990:3	–0.26	0.022	
M2	TC	1939:12	0.024	0.0055	War
	AO	1941:5	–0.037	0.0066	War
	LS	1941:6	0.011	0.0016	War
	TC	1944:12	–0.023	0.0055	War
	IO	1951:8	0.032	0.0070	Easier monetary policy
	TC	1951:10	0.027	0.0055	"Increase in confidence"
	IO	1952:4	–0.029	0.0070	
	IO	1952:6	0.028	0.0070	"Influx of deposits"
	LS	1952:7	–0.0085	0.0017	
	AO	1982:11	–0.034	0.0071	
	TC	1988:11	–0.026	0.0056	
	TC	1989:2	0.023	0.0056	
	AO	1990:1	0.067	0.0072	
	IO	1990:3	–0.066	0.0074	
	AO	1990:12	0.065	0.0071	
	AO	1991:1	–0.16	0.0072	
	AO	1991:5	0.039	0.0072	

continued

Table 1. Continued.

Series	Type	Date	Size	S.E.	Event
	TC	1992:1	–0.024	0.0056	
	TC	1992:10	0.025	0.0056	
	TC	1993:2	–0.26	0.0056	
Exchange rate	AO	1922:2	–9.27	1.71	
	AO	1922:4	11.01	1.70	
	TC	1922:5	–12.87	0.94	
	IO	1922:10	–8.87	1.70	
	AO	1922:11	–9.34	1.59	
	IO	1923:12	8.04	1.65	
	IO	1931:12	18.17	1.65	Out of the gold standard
	IO	1932:1	9.74	1.65	Out of the gold standard
	AO	1932:7	8.22	1.56	
	AO	1939:9	–21.85	1.56	Onset of war
	AO	1940:3	–8.30	1.64	War
	IO	1940:4	–12.76	1.67	War
	IO	1940:5	–9.56	1.71	War
	IO	1940:6	6.69	1.67	War
	AO	1942:8	–6.42	1.56	War
	IO	1945:6	45.98	1.65	End of war, devaluation
	IO	1945:8	35.44	1.65	Devaluation
	AO	1945:10	–5.05	1.57	Devaluation
	TC	1946:1	–4.29	0.86	
	TC	1947:3	–7.86	0.86	
	TC	1947:11	–5.05	0.86	
	TC	1949:7	6.01	0.85	Devaluation
	AO	1950:6	–7.20	1.56	
	TC[2]	1957:9	16.00	1.23	Devaluation
	IO[3]	1957:9	14.92	0.79	Devaluation
	AO[3]	1957:10	9.29	0.75	
	TC[3]	1957:11	–7.75	1.23	
	IO	1967:10	17.45	0.79	Devaluation
	AO	1968:1	–9.22	0.75	
	AO	1977:4	5.47	0.75	Devaluation
	IO	1982:10	7.32	0.79	Two devaluations
	TC	1991:11	4.66	0.69	Devaluation
	IO	1992:9	10.25	0.79	The Markka floats
	IO	1993:2	5.36	0.79	
	IO	1993:4	–4.55	0.79	Plan to cut public spending
	AO	1994:10	–3.67	0.75	Referendum to join the EU
Terms of trade	AO	1940:11	–5.57	1.19	War
	IO	1942:5	–5.92	1.43	War
	AO	1948:3	6.54	1.21	
	TC	1948:6	7.62	1.37	
	AO	1949:5	5.63	1.19	
	AO	1951:11	–4.84	1.20	Reductions in domestic prices
	AO	1952:1	6.58	1.25	
	AO	1952:4	26.30	1.24	
	IO	1952:10	15.38	1.43	
	IO	1953:1	8.32	1.43	
	IO	1953:7	–21.30	1.43	
	IO	1953:8	11.38	1.43	
	AO	1953:10	–10.33	1.19	
	AO[2]	1955:6	7.72	1.19	
	AO[2]	1958:12	8.48	1.19	
	TC[3]	1959:1	–5.17	1.18	
	AO[2]	1959:6	–5.68	1.19	
	AO	1967:12	–7.90	1.58	

continued

Table 1. Continued.

Series	Type	Date	Size	S.E.	Event
	IO	1973:3	9.09	1.60	
	IO	1973:5	7.66	1.60	
	AO	1973:6	–6.45	1.58	
	AO	1975:5	–8.64	1.58	
	TC	1986:2	5.30	1.18	Reduction in oil prices
Stock exchange index	TC	1945:3	0.21	0.035	War
	IO	1945:6	0.18	0.041	End of war
	AO	1945:7	0.24	0.039	End of war
	AO	1945:9	–0.36	0.039	
	AO	1945:11	–0.18	0.039	
	IO	1946:2	–0.25	0.041	
	IO	1946:5	0.17	0.041	
	IO	1946:7	–0.16	0.041	
	TC	1992:9	0.18	0.40	The Markka floats
Stock exchange turnover	IO	1931:10	1.37	0.35	Out of the gold standard
	IO	1939:12	–1.88	0.35	Onset of war
	IO	1940:1	1.60	0.35	War
	IO	1940:4	1.81	0.35	War
	IO	1985:7	–1.60	0.30	
	AO	1990:2	–1.16	0.27	Liberalization of exchange markets
	TC	1992:9	0.61	0.14	The Markka floats

[1] Combined results from two detections, periods 1922:2–1963:9 and 1955:2–1996:9; [2] Detected in the first period; [3] Detected in the second period.

outliers are given by series, with their type, timing, magnitude and standard error. Table 2 provides a chronological summary of the detected outliers. These results will be examined in some detail in this section. The smallest number of detected outliers per series (both subperiods) is five (bankruptcies), and the largest is 35 (exchange rate). The total number of outliers in all series is 187, of which 72 are AOs, 67 IOs, 6 LSs and 41 TCs, and in one case the same observation was classified as a TC in the first subperiod and an IO in the second subperiod.

The overall amount of outliers in this data is therefore roughly two per cent of all observations, ranging in individual series approximately from 0.6% to 4%. If outliers are considered rare events, these numbers are certainly appropriate, perhaps even somewhat conservative. It is easy to accept that one observation in one hundred is an outlier for any of these series. This implies, on average, an outlier approximately every eight years. Similarly, an outlier probability of five per cent for the more turbulent series, say the foreign exchange and price index series, is not unreasonable either (implying an outlier every other year, on average).

Perhaps surprisingly, there are practically no outliers at all in the period before the Second World War, apart from the wholesale price index and the exchange rate series.[1] This must partly be due to the fact that the Great Depression was not as severe in Finland as it was in many other countries. In comparison, several outliers are detected during and immediately after the Second World War, and into the late 1950s. After some years of few outliers, a larger cluster of them are again detected in the late

[1] One possible explanation for this would be the influence of the World War II on the results. However, if the outlier detection is performed only on the pre-WWII data, the results stay mostly the same. The exceptions are the BA and TT series, where additional outliers are indeed detected (4 and 10, respectively). On the other hand, in the case of the FX series, six previously detected outliers are not detected from the pre-war data. Therefore, the volatility associated with the war is not always the most influential on these series.

Table 2. Number of detected outliers in each year.

Year	IP	BA	CPI	WPI	LEND	M1	M2	FX	TT	SX	ST	Total
1922				1				5				6
1923			1					1				2
1924												
1925												
1926												
1927				1								1
1928												
1929												
1930												
1931				2				1			1	4
1932				1				2				3
1933												
1934												
1935												
1936												
1937												
1938												
1939	1	1		1		1	1	1			1	7
1940					2			4	1		2	9
1941	1	1		1	3		2					8
1942			1	1	2			1	1			6
1943		1			1							2
1944					3		1					4
1945		1	2	3				3		5		14
1946				1	1			1		3		6
1947			2	3				2				7
1948			1						2			3
1949								1	1			2
1950	2		1					1				4
1951				1			2		1			4
1952							3		3			6
1953									4			4
1954			1	1								2
1955	1								1			2
1956	2		2	1								5
1957				1	1			3				5
1958	2								1			3
1959	1								2			3
1960												
1961												
1962												
1963												
1964		1	1									2
1965												
1966												
1967	1		2	1				1	1			6
1968	1		2					1				4
1969												
1970	1											1
1971	2											2
1972			1									1
1973			1	1		2			3			7
1974			1	1								2
1975			1			1			1			3
1976				1								1
1977	1							1				2
1978						1						1
1979												
1980												

continued

Table 2. Continued.

Year	IP	BA	CPI	WPI	LEND	M1	M2	FX	TT	SX	ST	Total
1981												
1982							1	1				2
1983												
1984												
1985				1	1						1	3
1986									1			1
1987												
1988					1		1					2
1989						1	1					2
1990						2	3				1	6
1991			1		2		2	1				6
1992							2	1		1	1	5
1993					1		1	2				4
1994								1				1
1995					1							1
1996												
Total	16	5	21	24	19	8	20	35	23	9	7	187

1960s and early to mid 1970s. The last notable cluster of detected outliers occurs at the end of the 1980s and the first half of the 1990s.

The last column of Table 1 provides economic and political events that may underlie the detected outliers. It must be kept in mind that these are not necessarily the causes of the outliers. They should rather be taken as events that may have been partially responsible for the aberrant observations in the data. For many of the detected outliers such causes can not be found, however. In finding potential events to explain the detected outliers, Bank of Finland (Monthly) Bulletins have been used, in addition to Hjerppe (1989). The Second World War is obviously the most important event during the period, and several outliers are detected in most series during it. Finland was engaged in open hostilities first from November 1939 (from the last day of the month, so the effects can only be seen in December) to March 1940, and again from June 1941 to September 1944.

Since Balke and Fomby (1994) noted that more outliers occur at the turning points of the business cycle and during recessions, it is interesting to see whether a similar conclusion holds for the Finnish data as well. Unfortunately, the timing of the Finnish business cycles varies somewhat from author to author. For the years before 1980, one source is Hjerppe (1989), and for the 1980s and 1990s Takala and Virén (1997) has been used.

I will now examine the series one by one. The industrial production series has half of its 16 detected outliers in the 1950s. Somewhat curiously, neither the beginning of the fighting at the end of 1939, nor the effects of the peace treaty of spring 1940, when a large part of the country was ceded to the Soviet Union, show up as outliers. Similarly, the general strike of March 1956 is not detected as an outlier, although a positive outlier is detected in the following month. Other strikes are detected, however, in September 1950. Note that the variance of the series decreases considerably after the 1950s, which leads to the finding that all of the outliers detected during the overlap of the subperiods are detected only in the second period. Indeed, there are no more outliers in the series after 1977. The bankruptcies series has only five outliers, all of them negative, and four of which appear during the war years 1939 to 1945.

The price index series have 21 (CPI) and 24 (WPI) detected outliers. The post-war inflation continued to the end of the 1940s, and produced several large outliers in both series. Many of these peaks are related to the lifting of various wartime controls. Government actions have been important later as well, since food prices were regulated, and for example the negative outlier in CPI in November 1954 may be due to an administrative reduction in food prices. An AO in October 1957 is detected in WPI in both periods, with almost exactly the same estimated magnitude, but other outliers are again only detected in the second subperiod.

The monetary aggregate series M1 and M2 have 8 and 20 outliers, respectively. The M2 has five outliers in the beginning of the 1950s, which are probably related to an easing of monetary policy, an influx of deposits, and the paying off of debt by export firms. There is also a cluster of eight outliers (four positive, four negative) in the two years at the very beginning of the 1990s, indicating an increase in the volatility of the series. During the same period there are only two outliers in the M1 series. Of the bank lending series' 19 outliers, 11 occur during the war years, and six in the period from 1985 to 1995. In this latter period a liberalization of the Finnish financial markets was followed first by an unprecedented credit expansion and then by a banking crisis and a major recession. These events must also be related to the outliers detected in the monetary aggregate series. In these three series there is only one outlier detected in the overlapping period (an IO in bank lending), which is again detected only in the second period.

The exchange rate series has the most detected outliers in the data, a total of 35. These appear often in clusters, indicating ARCH type effects rather than isolated outliers. The first of these clusters consists of five outliers (three negative, two positive) in the ten months from February to November in 1922. The aftermath of leaving the gold standard at the end of 1931 shows up as two IOs in the next periods. There are also four consecutive outliers immediately after the truce of 1940. The most marked feature of the post-war period are the frequent devaluations, sometimes occurring twice a month, which often show up as outliers. Note, however, that not all of the minor devaluations are detected as outliers. The Finnish Markka was let to float in September 1992, resulting in some volatility in the following years. During the overlap, September 1957 is detected as an (positive) outlier in both subperiods, although classified as a TC in the first and an IO in the second (note that in practice there may not be much difference between an IO and a TC in an AR model). Moreover, there seems to be another outlier immediately afterwards. October is classified as a positive AO in the second period, and November as negative TC in the first. The results for the two subperiods are therefore quite similar.

The terms of trade series has a total of 23 outliers. There seems to be volatility in the series in the late 1940s and the first half of 1950s, where 15 outliers are detected. Overall, it is quite difficult to think of potential explanations for most of the outliers in the terms of trade series. Regarding the overlap, most outliers are only detected in the first subperiod, and the one outlier detected in the second does not quite agree with the results from the first.

Unexpectedly, the stock exchange series have relatively few outliers. The UNITAS stock exchange index has only nine outliers in all, eight in the years 1945 to 1946, and one in 1992. The turnover series has seven outliers, three of which occur during the war. The clusters of consecutive outliers could again be viewed as volatility rather than outliers. No outliers are detected in these series during the overlap.

As can be seen, for many outliers a reasonable underlying event can be found. Nevertheless, there are still a large number of observations where such events could not be easily found. Also, in some cases the timing of the business cycle does not seem to agree with the detected outlier (e.g. large negative outliers in industrial production during expansions). More detailed research into these problems is beyond the scope of this paper, however.

A surprise in the results, already alluded to, is that many a priori important events do not show up as detected outliers in any series. Such events are, for example, the general strike of March 1956, the stock market crashes of 1929 and 1987, and the oil price shocks of the 1970s. Some of these are visible in the plots of the time series, but apparently are not aberrant enough, compared to the AR(4) models, to be classified as outliers.

Another surprise is that the results from the two subperiods do not agree. Of the 18 outliers detected during the overlapping period, only two are detected in both subperiods (and in one of these the outlier is identified differently in the subperiods, namely as a TC in the first and an IO in the second). The most likely explanation is that the variances of the series usually differ in the two subperiods. This will also influence whether the outlier detection procedure will consider an observation of a certain absolute value an outlier or not. In addition, differences in the estimated AR parameters in some series may also contribute to this. With different parameter estimates, the results of outlier detection (which are dependent on the estimated model) may change as well.

The first finding of Balke and Fomby (1994), that more outliers are detected during recessions, does not initially seem to hold for this data. Indeed, the converse seems to be the case. However, considering the fact that the recessions in Finland have been relatively few and quite short-lived, especially after the Second World War, it can be said that a larger percentage of recession observations are outliers than of expansion observations. In addition, a large number of outliers are detected at the turning points of the business cycle, which agrees with the findings of Balke and Fomby. The second finding of Balke and Fomby, that the detected outliers often cluster both within and across series is clearly the case here as well. The third finding, that nominal and real series differ in the amount of outliers detected and their influence, on the other hand does not seem to hold for this data (since the industrial production series has quite a lot of detected outliers).

4. *The sensitivity of time series modeling*

In this section the sensitivity of time series modeling with respect to the outliers is considered. First, standard errors and normality coefficients (skewness and excess kurtosis) are computed from the AR residuals. Second, nonlinearity tests are calculated for the series. These analyses are done both for the original and the outlier corrected data to see how the evidence of non-normality and nonlinearity are altered by taking the outliers into account. The outlier corrected data is the outcome of an intervention analysis, where the previously detected outliers are taken into account in the model with the use of intervention (dummy) variables.

First of all, however, it could be noted that changes in the estimated AR parameters were usually not great after the outliers were taken into account. Nevertheless, in some series, where a large number of outliers were corrected, some or all parameters could change noticeably. The clearest example of such changes is the exchange rate series in the first period, where also the greatest number of outliers was detected. Without the outlier corrections the AR parameters were all insignificant, and close to zero (with absolute values all well below 0.1), but when the outliers were taken into account, all became significant at the 1% level, and their values grew noticeably (to between 0.13 and 0.27). It seems therefore, that even though outliers can bias the estimated autoregression parameter values, the amount of outliers in these series is mostly not large enough to cause serious biases in the AR parameter estimates, apart from a few exceptions.

Table 3 has the coefficients of skewness and (excess) kurtosis, along with the standard errors of the residuals. The first three columns on the left are for the original series, and the next three for the outlier corrected series. Almost all of the original series have significant skewness, and excess kurtosis is notable. Normality is therefore clearly rejected for all series. After the outlier correction these measures of non-normality are decreased, sometimes quite dramatically. Excess kurtosis still remains significant for all series even after the outliers are removed, whereas skewness becomes insignificantly different from zero in many cases. Curiously, for LEND1 and TT2 skewness increases after the outlier correction. The reduction in standard errors is also quite dramatic for some series, the most notable example being the over 50% decreases in the exchange rate series. This is of course mostly due to the removal of the largest observations as outliers. Jarque-Bera statistics

Table 3. AR(4) model residual statistics.

	Original data			Outlier corrected data			
Series	Skew.[1]	Kurt.[2]	SE[3]	Skew.[1]	Kurt.[2]	SE[3]	∇SE[4]
IP1	–0.59	2.62	0.069	–0.19	0.86	0.065	7.1%
IP2	–0.91	9.24	0.036	–0.27	1.29	0.026	28.6%
BA1	–0.47	3.13	0.36	0.12	1.53	0.34	7.6%
BA2	–0.21	1.35	0.21	–0.074	0.91	0.20	1.8%
CPI1	2.68	19.27	0.017	0.50	1.39	0.011	33.3%
CPI2	2.21	23.62	0.0065	0.70	0.81	0.0044	31.1%
WPI1	2.61	13.79	0.016	0.26	0.92	0.0090	42.8%
WPI2	1.50	5.63	0.0061	0.49	1.01	0.0051	16.1%
LEND1	–0.030	5.60	0.013	–0.15	0.92	0.0099	22.5%
LEND2	0.54	4.66	0.0061	0.038	0.81	0.0051	15.8%
MI1	0.69	2.98	0.023	0.23	0.53	0.022	4.6%
MI2	0.51	9.86	0.030	0.16	0.61	0.024	21.9%
M21	0.41	2.15	0.0082	–0.091	0.40	0.0066	15.0%
M22	–3.87	52.83	0.012	0.23	1.26	0.0073	38.6%
FX1	3.43	46.95	3.87	0.23	1.65	1.65	57.3%
FX2	5.06	47.78	1.59	–0.16	5.01	0.79	50.4%
TT1	0.65	31.37	2.49	0.037	1.63	1.43	42.6%
TT2	0.018	3.69	1.83	–0.023	1.64	1.60	12.6%
SX1	–0.27	7.12	0.050	0.036	1.08	0.041	19.5%
SX2	0.22	1.57	0.048	0.11	1.32	0.047	1.9%
ST1	0.20	2.92	0.38	0.0043	0.69	0.35	8.0%
ST2	–0.25	1.60	0.32	0.031	0.48	0.30	5.6%

[1] Skewness, standard error is 0.11. [2] Excess kurtosis, standard error is 0.22. [3] Standard error. [4] The decrease in SE due to outliers: (SE(original data) – SE(outlier corrected data))/SE(original data).

were also calculated for the series, but will not be reported since they are highly significant for almost all series in both periods, even after the outlier correction. The two exceptions are the outlier corrected M21 and ST2, which are not significant at the 5% level.

The nonlinearity tests that were computed will be presented and examined next. Lagrange multiplier (LM) tests for ARCH (Engle 1982) were computed with lags 4 and 8. The tests use an auxiliary regression of $\hat{\varepsilon}_t^2 = \alpha_0 + \alpha_1\hat{\varepsilon}_{t-1}^2 + \ldots + \alpha_i\hat{\varepsilon}_{t-i}^2 + \xi_t$, where $\hat{\varepsilon}_t$ are the estimated AR(4) model residuals, and $i = 4, 8$. The test statistic is TR^2, where T is the number of observations and R^2 is the multiple coefficient of correlation in the test regression, and it has an asymptotic χ^2 distribution (under the null hypothesis) with degrees of freedom i. Bilinearity is tested from two auxiliary regressions (see, e.g. Weiss 1986). The first, denoted BLN1, is that of $\hat{\varepsilon}_t = \gamma_0 + \gamma_1\hat{\varepsilon}_{t-1}x_{t-1} + \gamma_2\hat{\varepsilon}_{t-2}x_{t-2} + \gamma_3\hat{\varepsilon}_{t-3}x_{t-3} + \gamma_4\hat{\varepsilon}_{t-4}x_{t-4} + \gamma_5 x_{t-1} + \gamma_6 x_{t-2} + \gamma_7 x_{t-3} + \gamma_8 x_{t-4} + \xi_t$, and the second, denoted BLN2 is that of $\hat{\varepsilon}_t = \gamma_0 + \gamma_1\hat{\varepsilon}_{t-1}x_{t-1} + \gamma_2\hat{\varepsilon}_{t-1}x_{t-2} + \gamma_3\hat{\varepsilon}_{t-2}x_{t-2} + \gamma_4\hat{\varepsilon}_{t-1}x_{t-3} + \gamma_5\hat{\varepsilon}_{t-2}x_{t-3} + \gamma_6\hat{\varepsilon}_{t-3}x_{t-3} + \gamma_7 x_{t-1} + \gamma_8 x_{t-2} + \gamma_9 x_{t-3} + \gamma_{10}x_{t-4} + \xi_t$, where x_t are the series observations (either the original or the outlier corrected series). The test statistics of these tests are also TR^2, and they have asymptotic χ^2 distributions with degrees of freedom equal to the number of bilinear terms in the auxiliary regression, that is 4 and 6, respectively. Tests for smooth transition autoregression were also computed with two parameterizations. Both tests are based on an auxiliary regression of x_t on p lagged x_ts and nonlinear terms $x_t^{(p)}$. The test statistics are

$$\frac{(SSR_0 - SSR_1)/m}{SSR_1/(T-m-p-1)},$$

where SSR_0 is the sum of squared residuals from the AR model, SSR_1 is the sum of squared residuals from the auxiliary regression, m is the number of parameters in the auxiliary regres-

Table 4. Nonlinearity tests' *p* values – original data.

Series	ARCH4	ARCH8	BLN1	BLN2	LSTAR	ESTAR
IP1	<0.01	<0.01	0.44	0.45	0.32	0.99
IP2	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
BA1	<0.01	<0.01	<0.01	<0.01	0.016	0.72
BA2	0.011	0.020	0.025	0.14	0.01	0.51
CPI1	<0.01	<0.01	0.085	<0.01	<0.01	<0.01
CPI2	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
WPI1	<0.01	<0.01	<0.01	<0.01	<0.01	0.053
WPI2	0.024	0.022	0.039	0.035	0.011	0.66
LEND1	<0.01	<0.01	0.66	0.046	0.028	0.80
LEND2	0.033	0.17	<0.01	<0.01	<0.01	0.16
MI1	<0.01	<0.01	<0.01	<0.01	<0.01	0.044
MI2	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
M21	<0.01	<0.01	0.17	<0.01	<0.01	<0.01
M22	<0.01	0.030	<0.01	<0.01	<0.01	<0.01
FX1	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
FX2	0.070	0.34	<0.01	<0.01	<0.01	<0.01
TT1	0.048	<0.01	0.071	0.11	0.34	1.00
TT2	<0.01	<0.01	0.053	<0.01	<0.01	0.11
SX1	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
SX2	<0.01	<0.01	0.35	<0.01	<0.01	0.11
ST1	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
ST2	0.16	<0.01	0.10	0.090	0.014	0.70

sion, p the number of AR parameters, and T the sample size. The test statistics have an approximate F distribution, with degrees of freedom m and T–m–p–1. The first of these tests is against logistic smooth transition autoregression, denoted LSTAR. In this case the additional terms in the auxiliary regression are $x_t^{(p)} = x_t^{(p)L}$, which consists of the terms x_{t-i}^2, $i = 1, \ldots, p$ and $x_{t-i}x_{t-j}$, $i, j = 1, \ldots, p$, $i < j$. If, as in this case, $p = 4$, there are ten nonlinear terms in the auxiliary regression (i.e. $m = 10$). The second of these tests is denoted ESTAR, and is an optimal test against exponential smooth transition autoregression. In this case $x_t^{(p)} = x_t^{(p)E}$ has, for $p = 4$, $m = 26$, and consists of the terms x_{t-i}^2, $i = 1, \ldots, p$, $x_{t-i}x_{t-j}$, $i, j = 1, \ldots, p$, $i < j$ and $x_{t-i} x_{t-j}^2$, $i, j = 1, \ldots, p$. For the details of these tests, see Granger and Teräsvirta (1993) and van Dijk, Franses and Lucas (1999b). Note also that these two tests are similar in spirit to the well-known RESET test, and the LSTAR test is in fact identical to the general nonlinearity test proposed by Tsay (1986b). They are likely to have power against a wide variety of nonlinearities.

All these tests are computed for both subperiods, first for the original series with no outlier corrections, and then for the series where the effects of the detected outliers have been removed. The p values of the tests for the original data are given in Table 4. It should be kept in mind that the tests are here seen more as general tests for certain types of nonlinearity or model misspecification, rather than for ARCH, bilinearity and STAR models as such. They are likely to detect also other kinds of nonlinearity, and a significant test statistic for the ARCH test, say, is not necessarily due to ARCH effects in the data.

And not surprisingly, for the original data the ARCH tests' test statistics are mostly significant at the 5% level. At least one of the bilinearity tests' statistics is also significant, or very close to being so, for all series apart from the industrial production in the first period. The smooth transition autoregression tests LSTAR and ESTAR have also a large number of significant p values. The notable exceptions are industrial production and terms of trade in the first period. All in all, 88 test statistics are significant at the 1% level, 16 at the 5% level and a further 5 at the 10% level. There is at least some evidence of nonlinearity for all of the series (at

Table 5. Nonlinearity tests' p values – outlier corrected data.

Series	ARCH4	ARCH8	BLN1	BLN2	LSTAR	ESTAR
IP1	<0.01	<0.01	0.98	0.39	0.53	1.00
IP2	<0.01	<0.01	0.52	0.12	0.12	0.95
BA1	<0.01	<0.01	<0.01	<0.01	<0.01	0.13
BA2	0.012	<0.01	0.17	0.11	0.087	0.93
CPI1	<0.01	<0.01	<0.01	<0.01	0.020	0.75
CPI2	0.030	0.054	0.49	0.011	0.032	0.82
WPI1	0.20	0.097	0.022	<0.01	0.012	0.67
WPI2	<0.01	0.015	<0.01	<0.01	<0.01	<0.01
LEND1	0.039	0.25	0.34	<0.01	<0.01	0.56
LEND2	<0.01	<0.01	0.49	<0.01	<0.01	0.13
M11	<0.01	<0.01	0.18	<0.01	0.018	0.74
M12	0.72	0.83	0.064	0.23	0.23	0.99
M21	<0.01	<0.01	0.60	0.029	0.18	0.98
M22	<0.01	<0.01	0.15	0.30	0.069	0.91
FX1	<0.01	<0.01	0.27	0.58	0.28	0.99
FX2	0.014	0.016	0.29	0.051	0.081	0.93
TT1	<0.01	<0.01	0.12	0.28	0.63	1.00
TT2	<0.01	<0.01	0.011	0.089	0.086	0.93
SX1	<0.01	0.015	0.61	0.64	0.14	0.96
SX2	<0.01	<0.01	0.76	0.027	<0.01	0.39
ST1	0.21	0.14	0.044	0.19	0.12	0.95
ST2	0.26	0.014	0.14	0.072	0.10	0.94

least one of the tests is significant). All in all, over four fifths of the tests point to significant nonlinearity (83% of the tests have p values of less than 0.1). Together these results, and the ones on skewness and kurtosis given earlier imply that there are clearly features in the series which the AR model is incapable of capturing.

The results change, however, when the detected outliers are removed from the series. The tests' p values for the outlier corrected series are given in Table 5. Most changes occur in the bilinearity and STAR tests. Most often significant test statistics become insignificant when the outliers are taken into account, although there are also cases where the opposite takes place. Hidden nonlinearities, i.e. cases where there is more evidence of nonlinearity when outliers are removed, can be found in CPI1 for the BLN1 test, and in LEND2 and FX2 for the ARCH tests. Spurious nonlinearities, or cases where evidence of nonlinearity disappears when outliers are removed, can be found for the ARCH tests in WPI1, for BL and STAR tests in M2, FX and SX in both periods, and for all types of tests in M12 and ST1. In fact, there is hardly any evidence of nonlinearity in the outlier corrected M12 series.

The changes in the nonlinearity tests' p values do not seem to be related to the number of corrected outliers in the series in any way. It would seem likely that the more outliers are removed, the more changes there would be in the test results. This is not the case, however. For example, only four outliers were corrected in ST1, but the p values of all tests (except for BLN1) change from highly significant to clearly non-significant. These series are also plotted as an example in Figure 1. The upper panel of the figure contains the original data, and the lower panel the series where the detected outliers have been removed. Similarly, only one outlier was found in M11, but when it is taken into account the BLN1 and ESTAR tests' p values change from less than 0.01 and 0.044 to 0.18 and 0.74. On the other hand, 16 outliers were corrected in the TT1 series, but the test results remain essentially the same. Of the test statistics in Table 5, only 43 are significant at the 1% level, 17 at the 5% level and 10 at the 10% level. Compared with the results for the original data, the number of tests significant at

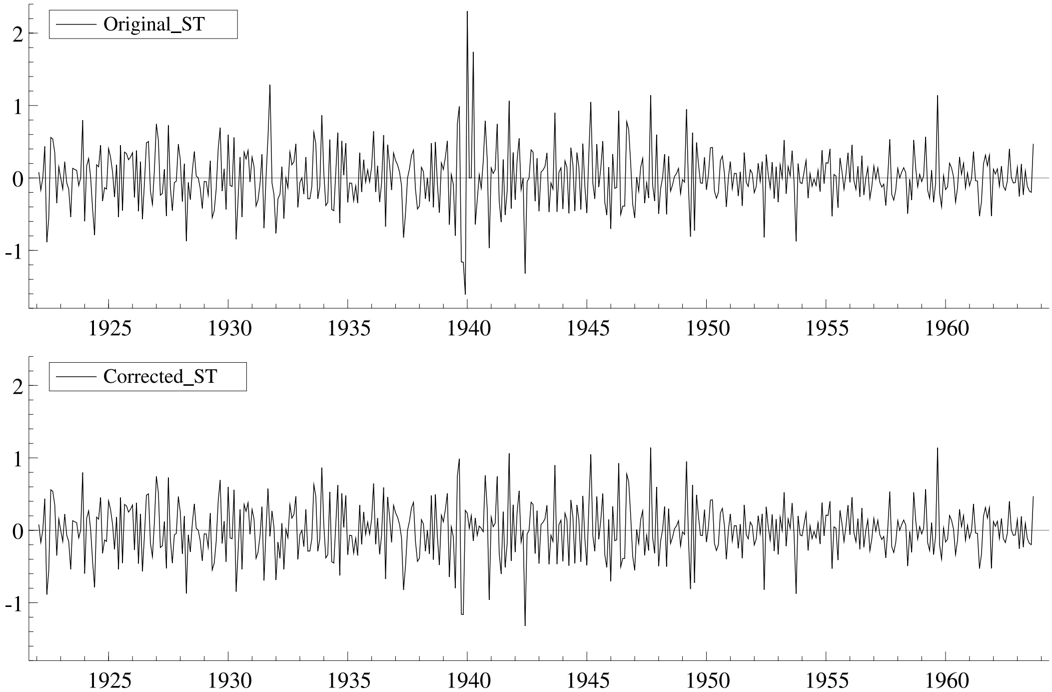

Figure 1. Stock market turnover, original and outlier corrected series.

the 1% level is reduced to less than half, whereas the number of tests significant at the 5 and 10% levels stays almost the same and doubles, respectively. Any evidence of nonlinearity is now found by just over half of the tests (53% of the tests have *p* values of less than 0.1).

On the other hand, where there is little doubt about nonlinearity in the data, the same conclusions can be drawn from both the original and the outlier corrected data. For example, exchange rate series are usually found to be nonlinear, and ARCH is a typically found feature. For this data the ARCH tests are highly significant in the first period, regardless of whether the outliers are taken into account. In the second period the rejection of the null hypothesis of no ARCH is more convincing once the outliers are removed from the data. The other nonlinearity tests on the other hand are no more statistically significant after the outlier correction. It should also be kept in mind that the outliers detected from the series were often clustered together, also indicating ARCH rather than isolated outliers. It is therefore reassuring to note that if there really is ARCH in the series (as we have good reason to expect in this case), then removing the detected outliers does not alter our inference about it.

To sum up these results, there is clear evidence of non-normality and nonlinearity in the series, both in their original form, and in their outlier corrected form. However, there is also a marked movement towards normality and linearity due to the outlier correction. Perhaps the most notable change is the decrease in skewness. Kurtosis remains significant even after the outlier correction, although it is also decreased a great deal. The changes in the results of the nonlinearity tests are somewhat mixed, although an overall decrease in the test statistics is quite clear.

Finally, a discussion on the role of the critical value in the outlier detection algorithm may be in order. Here a low sensitivity value, 4.0 for this sample size, was used. Balke and Fomby (1994) used a medium sensitivity value of 3.0. Note also that for longer series it is usually advisable to have somewhat larger critical values than for shorter series. Using a smaller, a more sensitive, critical value results in more detected outliers. At the same time the probability of labeling non-outlier observations as outliers obviously increases. For example, using medium and high sensitivity critical values of 3.5 and 3.0, respectively, a total of 6 and 23 outliers are detected in the SX2 series (compared with just one outlier found earlier with a critical value of 4.0). But as has been shown earlier, the changes in the nonlinearity tests' results are quite notable already with this amount of outliers. If a more detailed analysis were conducted, a more careful decision on the critical value should also be made, perhaps experimenting with different values, as suggested by Chen and Liu (1993). For the purpose of showing the potential effects of only a few outliers, however, the low sensitivity value used here seems adequate.

5. Discussion

I will first briefly repeat the main findings of this study. Several outliers are detected in the series, often during recessions and at the turning points of the business cycle. Outliers often cluster together both across series and within a series, and the price series have the most outliers (with the unexpected exception of stock market prices). The effects of the outliers on nonlinearity tests are notable, and can be in either direction, although in most cases evidence of nonlinearity is decreased when the outliers are taken into account. On the other hand, the effects do not seem to be dependent on the amount of detected outliers in the series. These findings mostly agree with those of Balke and Fomby (1994) on quarterly U.S. data.

The parameter estimates of the linear AR models seem to be fairly robust with respect to the effects of the detected outliers. On the other hand, the inferences from the nonlinearity tests can be highly dependent on only a few, or even a single observation. The same seems to hold also for the parameter estimates of nonlinear models. Although not reported here, GARCH(1,1) models were also estimated for the AR(4) models residuals. When the nonlinearity test results change after the outlier correction, the parameter estimates of the GARCH models seem to change as well. In several instances, the GARCH model could not even be estimated before the outliers were removed from the data. These findings may well apply also to other nonlinear models, although further research on this topic is obviously needed (but see Franses and van Dijk 1999 and Gabr 1998 for some results).

It is a well-known problem of nonlinear modeling that even with large data sets the nonlinear parameters may have to be estimated from only a small number of observations. Repeating the nonlinearity tests after the removal of detected outliers can be interpreted as a sensitivity analysis to see how robust the results are to minor variation in the data (explicitly robust tests are another possibility for this). And based on the results found here, the robustness of the nonlinearity tests is certainly questionable. There seems therefore to be scope for considering outlier models as an alternative to nonlinear models. Naturally, this is not appropriate in every case, and has also to depend on the ultimate aim of the modeling.

In any case, the message in the results of this paper is clear. In examining economic data any potential outliers should be taken seriously, no matter what the ultimate aim or the model being used may be. Outliers have already been shown to be potentially very harmful, and there is also increasing evidence that the dangers are not only theoretical.

References

Balke, N.S. (1993). "Detecting level shifts in time series." *Journal of Business & Economic Statistics* 11, 81–92.

Balke, N.S., and T.S. Fomby (1991). "Shifting trends, segmented trends, and infrequent permanent shocks." *Journal of Monetary Economics* 28, 61–85.

– **(1994).** "Large shocks, small shocks, and economic fluctuations: Outliers in macroeconomic time series." *Journal of Applied Econometrics* 9, 181–200.

Bank of Finland, Monthly Bulletin. Years 1921–1939 and 1946–1987. Bank of Finland, Helsinki.

Bank of Finland, Bulletin. Years 1988–1996. Bank of Finland, Helsinki.

Chen, C., and L.-M. Liu (1993). "Joint estimation of model parameters and outlier effects in time series." *Journal of the American Statistical Association* 88, 284–297.

Chen, C.W.S. (1997). "Detection of additive outliers in bilinear time series." *Computational Statistics and Data Analysis* 24, 283–294.

van Dijk, D., P.H. Franses, and A. Lucas (1999a). "Testing for ARCH in the presence of additive outliers." *Journal of Applied Econometrics* 14, 539–562.

– (1999b). "Testing for smooth transition nonlinearity in the presence of outliers." *Journal of Business & Economic Statistics* 17, 217–235.

Fischer, B., and C. Planas (1998). Large scale fitting of ARIMA models and stylised facts of economic time series. Eurostat working paper 9/1998/A/8. Eurostat, Luxembourg.

Franses, P.H., and D. van Dijk (1999). Outlier detection in the GARCH(1,1) model. Econometric Institute Research Report EI-9926/A. Erasmus University Rotterdam.

Franses, P.H., and N. Haldrup (1994). "The effects of additive outliers on tests for unit roots and cointegration." *Journal of Business & Economic Statistics* 12, 471–478.

Gabr, M.M. (1998). "Robust estimation of bilinear time series models." *Communications in Statistics – Theory and Methods* 27, 41–53.

Gómez, V., and A. Maravall (1994). Program TRAMO. Time series regression with ARIMA noise, missing observations, and outliers. Instructions for the user. EUI working paper 94/31. European University Institute, Florence, Italy.

– (1995). Programs TRAMO and SEATS. Update: December 1995. EUI working paper 95/46. European University Institute, Florence, Italy.

Granger, C.W.J., and T. Teräsvirta (1993). *Modelling nonlinear economic relationships.* Oxford: Oxford University Press.

Hjerppe, R. (1989). *The Finnish economy 1860–1985: Growth and structural change.* Helsinki: Bank of Finland.

Hoek, H., A. Lucas, and H.K. van Dijk (1995). "Classical and Bayesian aspects of robust unit root inference." *Journal of Econometrics* 69, 27–59.

Ledolter, J. (1989). "The effect of additive outliers on the forecasts from ARIMA models." *International Journal of Forecasting* 5, 231–240.

Sakata, S., and H. White (1998). "High breakdown point conditional dispersion estimation with application to S&P 500 daily returns volatility." *Econometrica* 66, 529–567.

Takala, K., and M. Virén (1997). Testing nonlinear dynamics, long memory and chaotic behaviour with financial and nonfinancial data. VATT discussion papers 147, Government Institute for Economic Research, Helsinki, Finland.

Tolvi, J. (2000). "The effects of outliers on two nonlinearity tests." *Communications in Statistics – Simulation and Computation* 29, 897–918.

Tsay, R. (1986a). "Time series model specification in the presence of outliers." *Journal of the American Statistical Association* 81, 132–141.

– (1986b). "Nonlinearity tests for time series." *Biometrika* 73, 461–466.

– (1988). "Outliers, level shifts and variance changes in time series." *Journal of Forecasting* 7, 1–20.

Weiss, A.A. (1986). "ARCH and bilinear time series models: Comparison and combination." *Journal of Business & Economic Statistics* 4, 59–70.