DISKUSSIONSBEITRÄGE

aus dem

Fachbereich

WIRTSCHAFTSWISSENSCHAFTEN

der

UNIVERSITÄT DUISBURG - ESSEN
Campus Essen

Nr. 148
Dezember 2005

The Temporal Disaggregation of Time Series

Andreas Kladroba

Andreas Kladroba

The Temporal Disaggregation of Time Series

1. Introduction

Most of the economic data is reported either quarterly or annually but it may happen that quarterly figures are required and only annual ones are available. Therefore, some different approaches for disaggregating annual data to quarterly data have been developed in the past years. In this paper we want to introduce some of the most popular methods (chapter 2) and then we want to compare them with the help of some simulations (chapters 3 and 4).

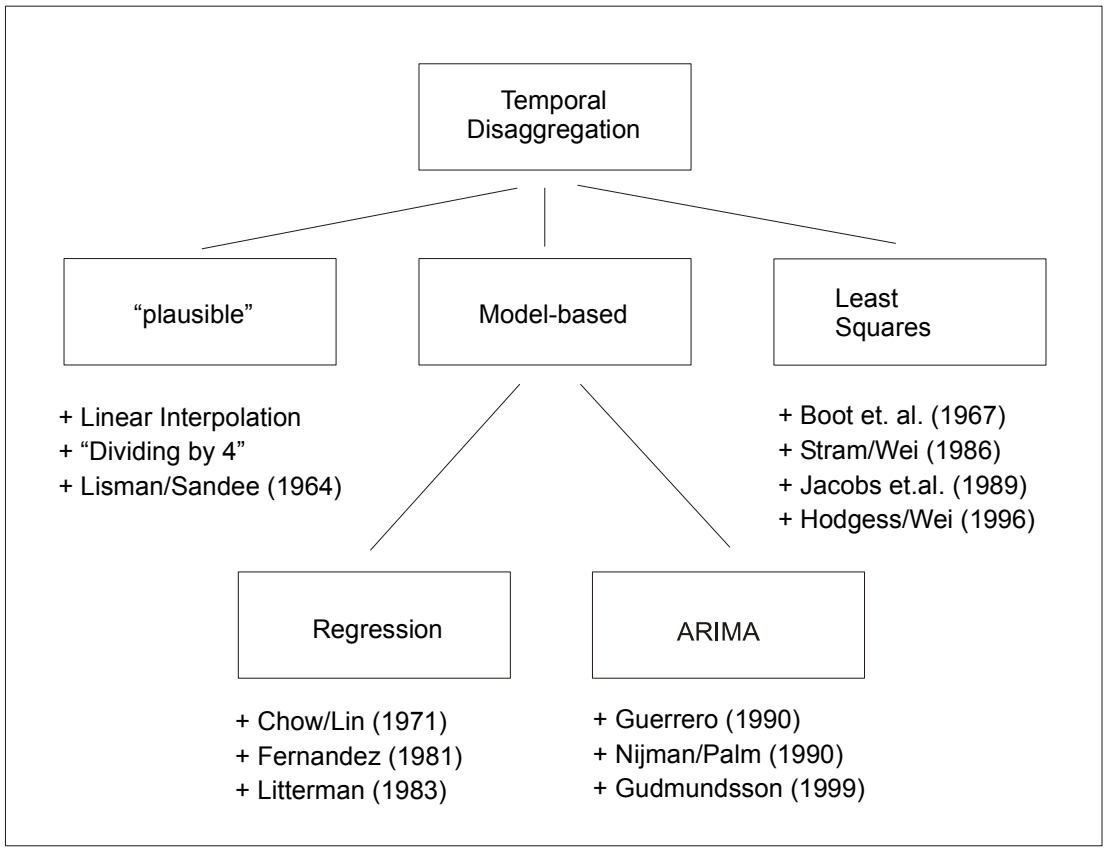

Fig. 1.1

According to Figure 1.1 the different approaches can be divided into the following categories:

1. In the first group we have methods which divide the annual data into a quarterly figure „in a plausible way“. This means linear interpolation if we have stocks and in the case of flows a simple „dividing by four“ of which the method of Lisman/Sandee is a special case.
2. The second group is formed by the so called model-based procedures. That means
   - first, using high correlating time series for creating the disaggregated series or
   - second, assuming that the wanted disaggregated series follow an ARIMA-process.
3. The third group („Least Squares“) tries to minimize the sum of the squared changes of the quarterly values respectively their d-th differences.

Assuming that y is the known T x 1 vector of the annual data for $t = 1,...,T$ and x is the unknown 4T x 1 vector of the quarterly data, then in case of aggregation we have the following connection:

(1.1) $y = C'x$

where C is a 4T x T aggregation matrix

(1.2) $C' = I_T \otimes e'$

while I_T is the T x T identity matrix and e is

(1.3a) $e = (1 \quad 1 \quad 1 \quad 1)$ in the case of flows

(1.3b) $e = (1 \quad 0 \quad 0 \quad 0)$ in the case of stocks if the first quarter is observed. Otherwise the „1“ has to be moved to one of the other positions.

Similary we get for disaggregation:

(1.4) $x = H'y$

with H being a T x 4T disaggregation matrix. In the following we have to show what the matrix H has to look like for the different procedures of disaggregation.

2. Some disaggregation procedures

2.1 „Plausible“ Methods

2.1.1 „Dividing by four“ and linear Interpolation

The first procedure we want to present is the simple „dividing by 4“ method in the case of flows. It is easy to see that the disaggregation matrix must be

(2.1) $$H' = \frac{1}{4}C$$

In the case of flows it must be considered that after disaggregation we only have 4(T-1) observations. We get the disaggregation matrix:

(2.2) $$H' = A + \frac{1}{4}BD$$

with $$A = \left[C'_{T-1} \quad 0_{4(T-1)}\right],$$

where $0_{4(T-1)}$ is a $4(T-1)$-zero vector. Furthermore, it is

$$B = \begin{bmatrix} \tilde{B} & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \ddots & \vdots \\ & & & 0 \\ 0 & & \dots\ 0 & \tilde{B} \end{bmatrix}$$ a 4(T-1) x 4(T-1) block matrix

with $$\tilde{B} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$ and

$$D = \begin{bmatrix} \tilde{D} & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \ddots & \vdots \\ & & & 0 \\ 0 & & \dots\ 0 & \tilde{D} \end{bmatrix}$$ a 4(T-1) x T block matrix with: $$\tilde{D} = \begin{bmatrix} -1 & 1 \\ -1 & 1 \\ -1 & 1 \\ -1 & 1 \end{bmatrix}.$$

2.1.2 The Procedure of Lisman/Sandee

In the case of the „dividing by four“ method at the beginning of every year there is a „step“ in the disaggregated time series. Lisman/Sandee (1964) want to avoid this by building a weighted mean of the quarterly values of the years t-1, t and t+1. So the procedure of Lisman/Sandee includes two steps.

In the first step we build the quarters of y_t:

$$\psi_{1t} = \psi_{2t} = \psi_{3t} = \psi_{4t} = \frac{1}{4} y_t \tag{2.3}$$

In the second step we get the weighted arithmetic means of $\psi_{i,t-1}$, ψ_{it} and $\psi_{i,t+1}$ for estimating x_{it}:

$$\begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{4,(T-1)} \end{bmatrix} = \begin{bmatrix} g_1 & g_2 & g_3 & 0_4 & 0_4 & 0_4 & \dots & 0_4 \\ 0_4 & g_1 & g_2 & g_3 & 0_4 & 0_4 & \dots & 0_4 \\ 0_4 & 0_4 & g_1 & g_2 & g_3 & 0_4 & & \vdots \\ 0_4 & 0_4 & 0_4 & g_1 & g_2 & \ddots & & \\ \vdots & \vdots & & & \ddots & \ddots & & 0_4 \\ 0_4 & 0_4 & \dots & & 0_4 & g_1 & g_2 & g_3 \end{bmatrix} \begin{bmatrix} \psi_{11} \\ \psi_{12} \\ \vdots \\ \psi_{4,T} \end{bmatrix} \tag{2.4}$$

with: $$g_1 = \begin{bmatrix} a & 0 & 0 & 0 \\ 0 & b & 0 & 0 \\ 0 & 0 & c & 0 \\ 0 & 0 & 0 & d \end{bmatrix}, \; g_2 = \begin{bmatrix} e & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & f & 0 \\ 0 & 0 & 0 & e \end{bmatrix}, \; g_3 = \begin{bmatrix} d & 0 & 0 & 0 \\ 0 & c & 0 & 0 \\ 0 & 0 & b & 0 \\ 0 & 0 & 0 & a \end{bmatrix}$$

respectively

$$x = \frac{1}{4} GCy \tag{2.4a}$$

Therefore we have a system of equations with six unknown Variables a,...,f,.which can be solved with the help of the following considerations:

(1) For all $t = 1,\dots,T$ it is

$$\sum_{i=1}^{4} x_{it} = 4\psi_t \tag{2.5}$$

So we have:

$$a + b + c + d = 0 \tag{2.6}$$

$$2(e + f) = 4 \tag{2.7}$$

(2) In the special case where $\psi_{t-1} = \psi_t = \psi_{t+1}$ we have:

(2.8) $a + e + d = 1$

(2.9) $b + c + f = 1$

(3) In the case of a rise/decline of the annual value by p we assume that the quarterly value in-/decreases by $\frac{1}{4}p$:

(2.10) $x_{it} - x_{i-1,t} = \frac{1}{4}p$

For the transition from the first to the second quarter, that is:

(2.11) $x_{2t} - x_{1t} = \frac{1}{4}p = (b-a)\psi_{t-1} + (f-e)\psi_t + (c-d)\psi_{t+1}$

Because of $y_t - y_{t-1} = p \Rightarrow \frac{y_{t-1}}{4} = \psi_{t-1} = \frac{y_t - p}{4} = \psi_t - \frac{p}{4}$ and $\psi_{t+1} = \psi_t + \frac{p}{4}$ we get from (2.10):

(2.12) $\frac{p}{4} = (b - a + f - e + c - d)\psi_t + (a - b + c - d)\frac{p}{4}$

and because of (2.5) and (2.6) we get

(2.13) $a - b + c - d = 1$

In the same way for the transition from the second to the third quarter we get:

(2.14) $2(b - c) = 1$

So we get a system of six equations of which only five are independent. The solution depends on an unknown variable α:

$a = \frac{3-\alpha}{4}$ $b = \frac{1+\alpha}{4}$ $c = \frac{\alpha-1}{4}$ $d = -\frac{3+\alpha}{4}$ $e = 1+\frac{\alpha}{2}$ $f = 1-\frac{\alpha}{2}$

For determining α, Lisman/Sandee propose using prior information like a known season figure.

2.2 Modelbased Procedures

2.2.1 Regression model-based Procedures

Assuming that $m \geq 1$ time series exist which are highly correlated with the wanted disaggregated series. We have

(2.15) $x = Z\beta + u$

with: Z: (4T x m) matrix of the correlating time series

β: (m x 1) vector of coefficients

u: (4T x 1) white noice

An unbiased estimator of x satisfies the requirements

(2.16) $\hat{x} = \hat{H}'y = \hat{H}'\left(C'Z\beta + C'u\right)$

So we have the expected estimation error

(2.17) $E(\hat{x} - x) = E\left[\hat{H}'\left(C'Z\beta + C'u\right) - (Z\beta + u)\right] = \left(\hat{H}'C'Z - Z\right)\beta = 0$

For an unbiased estimation the following must be valid:

(2.18) $\hat{H}'C'Z - Z = 0$

$\hat{x} - x = \hat{H}'C'u - u$

with the covariance

(2.19) $Cov(\hat{x} - x) = E\left[\left(\hat{H}'C'u - u\right)\left(\hat{H}'C'u - u\right)'\right] = \hat{H}'C'VC\hat{H} - \hat{H}'C'V - VC\hat{H} + V$

By minimizing this term we get the unbiased minimum variance estimator:

(2.20) $\hat{x} = Z\hat{\beta} + \left[C'VC\left(C'VC\right)^{-1}\right]C\hat{u} = Z\hat{\beta} + C\hat{u}$,

where $\hat{\beta}$ is the GLS-estimator using the T aggregated data and $\hat{u}$ the corresponding residual vector:

(2.21) $\hat{\beta} = \left[Z'C\left(C'VC\right)^{-1}C'Z\right]^{-1}Z'C\left(C'VC\right)^{-1}Cy$

(2.22) $C\hat{u} = y - C'Z\hat{\beta}$

2.2.2 ARIMA-based Models

Assuming that the wanted disaggregated time series follows an ARIMA(p,d,q)-process:

(2.23) $\phi(B)(1-B)x_t = \tau(B)\varepsilon_t$,

where B is the shiftoperator $Bx_t = x_{t-1}$ and ε_t is gaussian. In a similar way as in chapter 2.2.1 based on the conditional mean of x

(2.24) $E(x_t|x_1, x_2, ...) = E(x_t)$

we get the unbiased minimum variance estimator:

(2.25) $\hat{x} = E(x) + \theta\theta' C(C'\theta\theta' C)^{-1}[y - C'E(x)]$

with the estimation error

(2.26) $x_t - E(x_t) = \sum_{j=0}^{t-1} \theta_j \varepsilon_{t-j}$

where θ_1, θ_2, ... is the solution of

(2.27) $\theta(B)\phi(B)d(B)\tau^{-1}(B) = 1.$

2.3 Least Squares Models

The last group we want to look at is formed by the models of Least Squares Estimation. We start by building the (4T – d) vector of the d-th differences of x.

(2.28) $w = \Delta_{4T}^d x$

with: $\Delta_l^d = \begin{bmatrix} \delta_0 & \delta_1 & \dots & \delta_d & 0 & \dots & & 0 \\ 0 & \delta_0 & \delta_1 & \dots & \delta_d & 0 & \dots & 0 \\ & & & \vdots & & & & \\ 0 & \dots & & & 0 & \delta_0 & \delta_1 & \dots & \delta_d \end{bmatrix}$ (l-d) x l

where the δ_i are the coefficients of the B_i in $(B-1)^d$. In the same way we build a vector u with the d-th differences of y.

We use the GLS-approach for estimating the disaggregated time series

(2.29) $\min_x w' V_w^{-1} w$,

where V_w is the covariance matrix of w.

The solution of this adjustment problem contains two steps:

1. Estimating w based on u.
2. Estimating x based on w.

Between u and w there is the following connection:

(2.30) $u = \Xi^d w$

with: $\Xi^d = \begin{bmatrix} \xi' & 0 & \dots & & 0 \\ 0_4 & \xi' & 0 & \dots & \vdots \\ 0_8 & & \xi' & & \\ \vdots & & & \ddots & \\ 0_{4(T-d-1)} & \dots & & & \xi' \end{bmatrix}$ (T – d) x (4T – d).

ξ' is a (3d + 4) x 1 vector with the coefficients of the B^i in $\left(1 + B + B^2 + B^3\right)^{d+1}$ and 0_l are 1 x l vectors of zero.

So as an estimator of w we get:

(2.31) $\hat{w} = V_w \left(\Xi^d\right)' V_u^{-1} u = V_w \left(\Xi^d\right)' V_u^{-1} \Delta_T^d y$

and as an estimator of x

(2.32) $\hat{x} = \begin{bmatrix} \Delta_{4T}^d \\ \hline 0 \mid I_d \otimes e_4 \end{bmatrix}^{-1} \begin{bmatrix} V_w \left(\Xi^d\right)' V_u^{-1} \Delta_T^d \\ \hline 0 \mid I_d \end{bmatrix} y$

with: I_d: d x d identity matrix

e_4: 1 x 4 vector [1, 1, 1, 1]

3. Simulation

After describing all the procedures the next question has to be: Which one is the best? This question is hard to answer. Some of the methods are „optimal“ by their nature because they are built as the result of a consideration of optimization. For example the ARIMA- or the regression-based methods are estimators with a minimal

variance or the Least Square method results from the minimization of the squared differences of the estimated quarterly values. Otherwise methods like Lisman/Sandee or „dividing by four“ do not originate from such optimization considerations. Are the first ones therefore better than the second ones? The most important question for getting the „best“ method is: Which procedure delivers an estimation closest to the original disaggregated time series? To answer this question it seems to be useful to carry out some simulations for comparing the different methods. We did several of them with different time series. In the following, we want to describe the building and the results of one of them in detail. In chapter 4 we will shortly agree of the other ones.

The simulation we want to describe is based on an ARIMA(1,1,1)-process as the disaggregated time series of a flow. For all calculations we have created „optimal“ conditions. This means that

- in the case of Lisman/Sandee several α were tried out and the one that delivered the best results was used
- the correlations of the reference series which were used in the regression model-based estimations were $\geq |0.95|$
- the „real“ coefficients were used for the estimation of the ARIMA-based model
- the variances V_w and V_u that were needed for the Least Squares model were calculated from the original time series

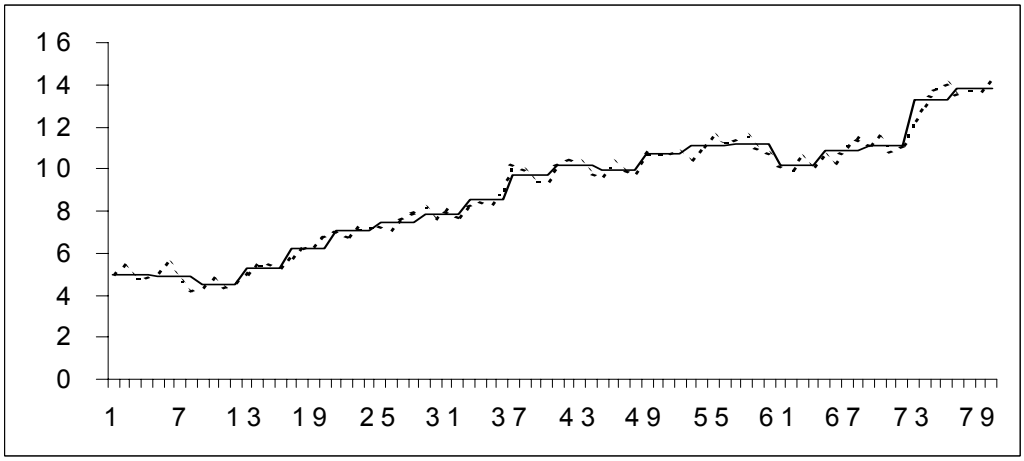

Fig. 3.1: Dividing by 4[1]

[1] In the figures 3.1 – 3.5 the full line represents the estimated time series and the pointed line represents the original time series.

The estimations of the disaggregated time series amount to the following results:

- The simple „dividing by 4“ method comes to a relatively good adaption with a fairley small MSE (fig. 3.1). The procedure of Lisman/Sandee (fig. 3.2) amount to the worst result of all with a fairely high MSE. It can be decreased by the choice of other – theoretically not to be explained and not to be reached by varying α - weights.

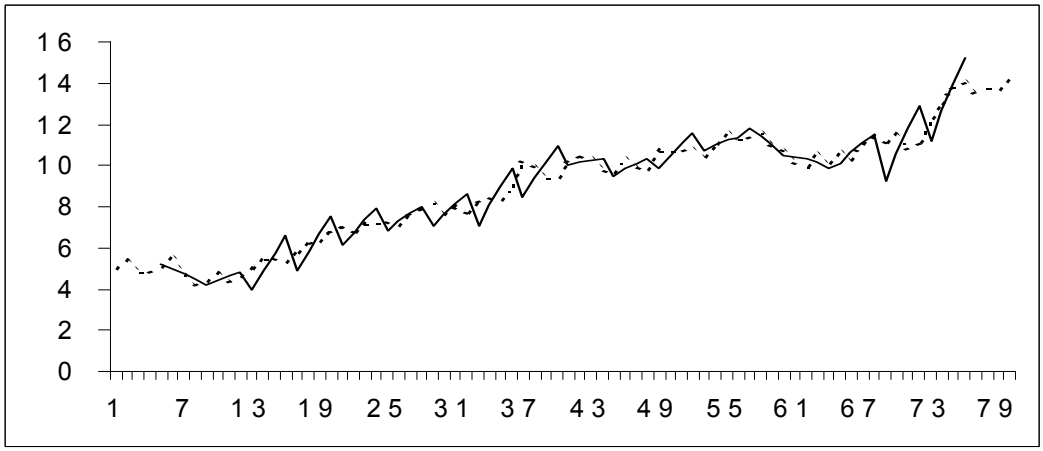

Fig. 3.2: Lisman/Sandee

- The regression model-based procedure (fig. 3.3) leads to the best result but it must be considered that the estimation very strongly depends on the correlation between the original and the reference series. A smaller correlation leads to worse results.

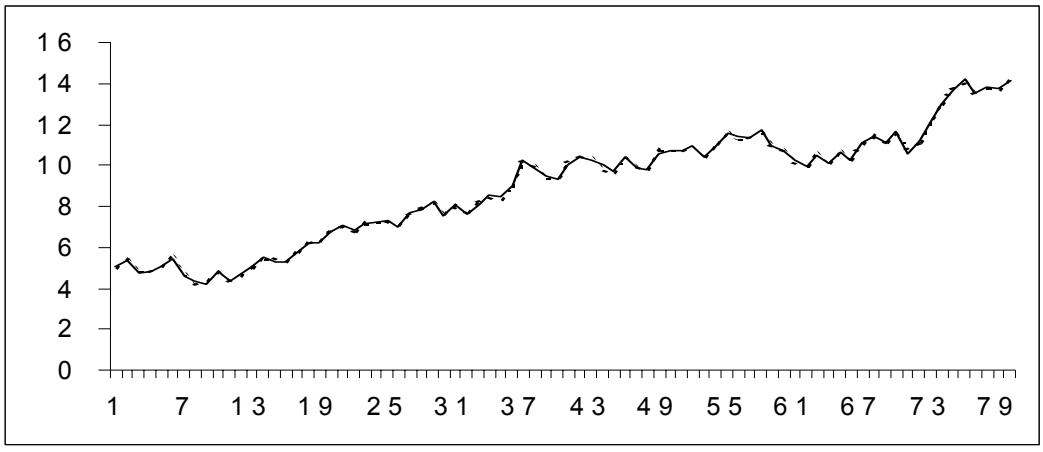

Fig. 3.3 Regression model-based estimation

- The ARIMA-model (fig. 3.4) comes to a useful MSE, too. However the results were worse than the results of the regression-based estimation. Similar to the regression-based model the estimation is very sensitive to a false assumption of the ARIMA structure.

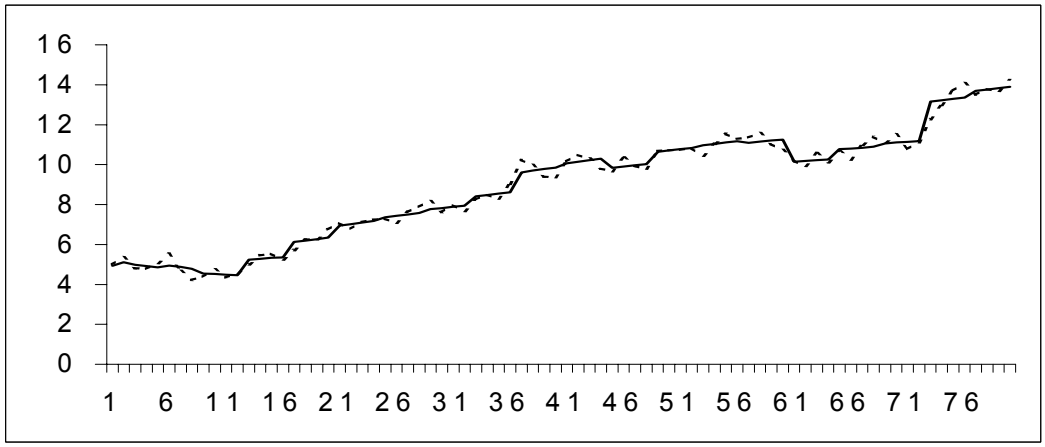

Fig. 3.4: ARIMA

- The results of the Least Squares model (fig. 3.5) turned out to be worse than those of the modelbased estimation. First of all the MSE is higher and second the sum of the estimated quarterly values is not even equal to the observed annual value. Moreover the estimation reacts very sensitiv to a false estimation of the covariances of w and u.

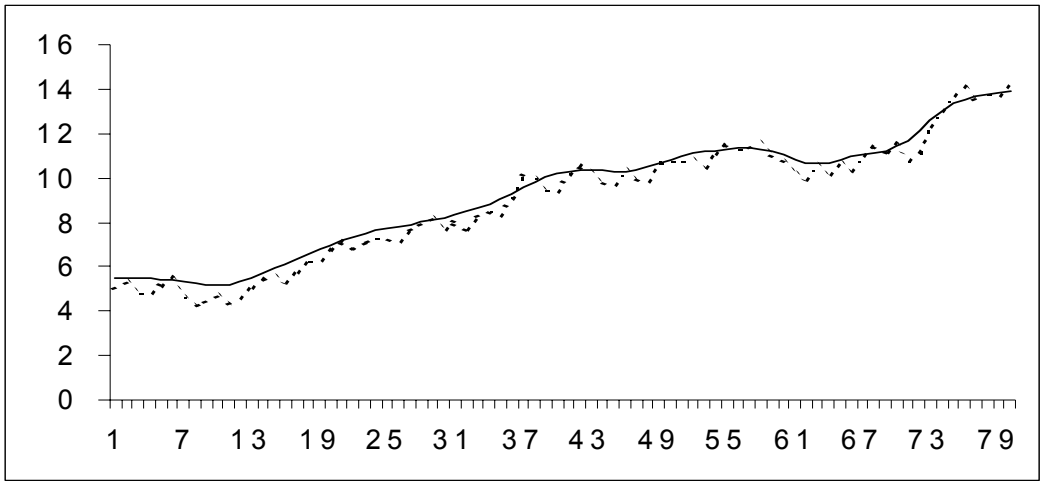

Fig. 3.5 Least Squares

4. Conclusion

In chapter 3 we mentioned that we did more than just the one simulation we described. These ones were based on a MA(2), on an AR(2), and on an ARMA(1, 1) process. The results of these simulations were similarly to the one of the ARIMA process. The most important difference is that the not model-based methods (dividing by four, Lisman/Sandee and Least Squares) delivered estimations with a smaller

MSE in the cases of not integrated time series compared to the one of the ARIMA process. But here the results also fall behind those of the model-based procedures.

Even after doing the simulations it is hard to answer the question: Which method is the best of all? Based on our criterion „precision of the estimation“ (measured by the MSE) it seems that the model-based methods deliver the better results. But also the simple „dividing by four“ comes to a low MSE. Here, we see an important difference between the two methods. The good results of the model-based procedure can only be reached if we can guarantee „optimal conditions“. In the other case the MSE can be much higher. Such a restriction does not exist for the „dividing by four“ method. So we have a case of „uncertainty“. The same problem exists for the Lisman/Sandee procedure (estimation of α), for the ARIMA-based method (finding the ARIMA structure of the reference series) and the Least Squares model (setting covariance). As a third criterion for choosing one of the methods we should pay attention to the expenditure of the estimation.

	Dividing	Lisman/ Sandee	Regression	ARIMA	Least Squares
MSE	low (8,55)	high (33,86)	very low (0,57)	low (7,57)	high (18,74)
expenditure	very low	low	middle	high	very high
uncertainty	none	high	high	high	very high

Tab. 4.1

Based on these criterions (precision, expenditure and uncertainty) we can sum up as follows (compare tab. 4.1):

1. Least Squares do not seem to be suitable for this application. The MSE of the described simulation but also of the ones not described is relatively high. Moreover the uncertainty and the expenditure are unreasonably highly.
2. It also seems to be better to avoid the application of the method of Lisman/Sandee. The MSE and the uncertainty are too high.

3. The model-based methods come to rather good results when conditions can be guaranteed. Then the relatively high expenditure can be justified.
4. Probably surprising are the good results of the „dividing by four“ method. Moreover the low expenditure and the absent uncertainty have to be acknowledged.

References

[1] Boot, J./W. Feibes/J. Lisman (1967), Further Methods of Derivation of Quarterly Figures from Annual Data, in: Applied Statistics 16, S. 65 – 75

[2] Chow, G/A. Lin (1971), Best Linear Unbiased Interpolation, Distribution, and Extrapolation of Time Series by Related Series, in: Review of Economics and Statistics 53, S. 373 – 375

[3] Fernandez, R. (1981), A Methodological Note on the Estimation of Time Series, in: Review of Economics and Statistics 63, S. 471 - 476

[4] Gudmundsson, G. (1999), Disaggregation of Annual Flow Data with Multiplicative Trends, in: Journal of Forecasting 18, S. 33 - 37

[5] Guerrero, V. (1990), Temporal Disaggregation of time Series: An ARIMA-based Approach, in: International Statistical Review 58, S. 29 – 46

[6] Hodgess, E./W. Wei (1996), Temporal Disaggregation of Time Series, in: M. Ahsanullah (Hrsg.), Applied Statistical Science I, Commack

[7] Jacobs, J./S. Kroonenberg/T. Wansbeek (1989), Dividing by 4: An efficient Algorithm for the Optimal Disaggregation of Annual Data into Quarterly Data, CCSO Series No. 7, Rijksuniversiteit Groningen

[8] Lisman, J./J. Sandee (1964), Derivation of Quarterly Figures from Annual Data, in: Applied Statistics 13, S. 87 – 90

[9] Litterman, R. (1983), A Random Walk, Markov Model for the Distribution of Time Series, in: Journal of Business & Economic Statistics 1, S. 169 – 173

[10] Nijman, T./F. Palm (1990), Predictive Accuracy Gain from Disaggregate Sampling in ARIMA Models, in: Journal of Business & Statistics 8, S. 405 - 415

[11] Stram, D./W. Wei (1986), A Methodological Note on the Disaggregation of Time Series Totals, in: Journal of Time Series Analysis 7, S. 293 - 302