Müller, Christian

Working Paper

On the effects of aggregating cointegrated variables over time

SFB 373 Discussion Paper, No. 2002,9

Provided in Cooperation with:
Collaborative Research Center 373: Quantification and Simulation of Economic Processes, Humboldt University Berlin

Suggested Citation: Müller, Christian (2002) : On the effects of aggregating cointegrated variables over time, SFB 373 Discussion Paper, No. 2002,9, Humboldt University of Berlin, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes, Berlin, https://nbn-resolving.de/urn:nbn:de:kobv:11-10048583

This Version is available at:
https://hdl.handle.net/10419/65360

On the Effects of Aggregating Cointegrated Variables over Time *

Christian Müller

Institut für Statistik und Ökonometrie
Wirtschaftswissenschaftliche Fakultät
Sonderforschungsbereich 373
Humboldt–University
Spandauer Str. 1
D-10178 Berlin, GERMANY
Tel.: +49-30-2093-5717
Fax: +49-30-2093-5712
Email: cmueller@wiwi.hu-berlin.de

This version: February 5, 2002

Abstract

It has long been recognized that aggregating time series introduces correlation between consecutive values of the aggregated observations (see Working (1960)). This paper investigates the effect of aggregation on the relation *between* variables assuming that the data generating process involves two integrated variables linked by a specific error correction mechanism (cointegration). It will be shown that aggregation does not distort the cointegration relation while some other features of the data generating process will change considerably. Cointegration tests become invalid in a single equation framework but system cointegration analysis seems to be robust against various aggregation strategies.

JEL classification: C32, C43

Keywords: cointegration, aggregation, time series

*I thank Rolf Tschernig and participants at the HELENENAU-Seminar 2001 for many helpful comments. Financial support from the Deutsche Forschungsgemeinschaft, SFB 373 is gratefully acknowledged.

1 Introduction

High frequency data such as interest rates or share prices are often aggregated over time, that is averaged, to give an impression of the general stance of the variable in a certain period. This is frequently used as a convenient tool for getting a grip on the overall performance. Sometimes these variables are also viewed in relation to other, less frequently measured macro-economic indicators such as the level of prices and the income of an economy. In these cases averages serve to represent the whole period for which the other variable is calculated once. In fact, most macro-economic data bases do not even provide the original, say daily data on interest rates rather than monthly or quarterly aggregates. Consequently these aggregates are used for statistical analysis simply for practical reasons.

The aggregation has its price, though. We will investigate the costs of aggregation in the context of cointegrated variables where the adjustment to the long-run relation is very fast. The motive for this setup is the frequently observed difficulty to establish a long-run relation between long-term and short-term interest rates despite distinguished theoretical arguments.[1] There we have both aggregated variables and (theoretically) very fast adjustment to the long-run relation whose parameters are also known.

Despite the early article by Working (1960) and the widespread use of aggregated data in empirical analysis, the hassles incurred by aggregation have not attracted much attention. The more recent contribution by Hallerbach (2000) points out the effect on the relationship between variables and for the first time suggests a filter for coping with the moving average component introduced by averaging the data. This paper extends the analysis in that it considers cointegrated variables and different aggregation strategies.

A distinction has to be drawn to other approaches such as Lütkepohl (1993) and the remarks made by Hamilton (1994).[2] There, stable, general *VARMA(p,q)* are considered and the interest focus on whether or not the resulting data series also

[1] see e.g. Gottschalk (1999), Müller & Hahn (2001).

[2] See pages 230ff in Lütkepohl (1993) and pp 106f in Hamilton (1994).

posses any valid *VARMA* representation. Here, non-stationary series are considered whose stationary error correction representation is not strictly *VARMA*. It is to be seen, that this difference will also effect the results. Moreover, we are not interested in whether or not the resulting series also have *any* error correction representation rather than whether or not they feature exactly the same error correction parameters as the underlying true process. This is because the primary interest is with the economically meaningful relationships between the variables which are to be disclosed. Therefore, the question to be answered could be put as: Is it possible to derive the economically interesting relationships between dis-aggregated data (the economic relations between the variables) from the observable aggregated data series?

Along the way to the answer to that question a few more information about the changing features of the DGP will be provided and for some of the more interesting details, a Monte Carlo study will provide an illustration.

The structure of the paper is as follows. First, we give a definition of the data generating processes and the aggregation methods we want to investigate. After that, the effects of aggregation on the aggregated series are calculated providing several hypotheses for a small simulation study which precedes the conclusions.

2 Data Generating Process and Aggregation Methods

2.1 A Brief View Over Existing Results

Apart from the notion of preserving the *cointegratedness* of aggregated cointegrated processes to be found e.g. in Hamilton (1994), researchers have also addressed a number of related questions among them for example forecasting performance Tiao (1972), seasonal and zero frequency unit roots Granger & Siklos (1995), Granger and instantaneous causality Mamingi (1996) and Breitung & Swanson (1998), cross-equation correlation and their filtering Hallerbach (2000). For many of these, one of the goals was to give advice on what aggregation method to prefer.

Table 1: Surveying Selected Articles on Aggregation of Time Series

Author	DGP	Aggregation	Topic	Conclusion
Working (1960)	univariate ARMA	averaging	Representation, error properties	averaging introduces MA effects in Residuals
Tiao (1972)	univariate ARMA	averaging	forecasting, asymptotic properties	forecast horizon matters; stationary processes: aggregation no problem, nonstationary processes: use original series
Granger & Siklos (1995)	uni- and multivariate ARMA	temporal averaging, systematic sampling	seasonal unit roots aliasing frequencies	systematic sampling may lead to spurious cointegration
Mamingi (1996)	bi-variate VECM	temporal averaging, systematic sampling	Granger causality	temporal aggregation distorts more heavily Granger causality results
Breitung & Swanson (1998)	uni- and multivariate (non-) stationary, cointegrated	temporal averaging, systematic sampling	asymptotics for causality; Granger and contemporaneous, selecting the correct causality ordering	temporal ordering of cause and effect mixed up, identification of contemporaneous causality impossible
Hallerbach (2000)	stationary univariate (pairs)	averaging, base series	auto- and cross correlation between base series, aggregated series, estimation biases filtering unwanted effects	cross correlation; no difference if averaging both series, reduction if one series averaged only
this study	cointegrated VAR	temporal averaging, systematic sampling	DGP representation, estimation bias, cointegration test	cointegration relations remain, other features of the DGP change considerably: correlation of error terms and regressors. Systematic sampling of averaged time series eases some spurious effects. Multivariate analysis tools preferable.

Granger & Siklos (1995) for instance recommend not to use systematic sampling (which will be referred to as skip sampling below) when seasonally varying data is involved because that may shift the seasonal unit root to the zero frequency and thereby introduce spurious cointegration under certain conditions. On the other hand Mamingi (1996) concludes on the basis of Monte Carlo Simulations that skip sampling might be superior when Granger causality is what matters most because this aggregation method features fewer distortions of related tests. To some extent his work is most closely related to the current one due to the similarities in the data generating processes employed and some of the conclusions drawn.

The present paper reaches beyond the objectives of the aforementioned work however, in that it treats the problem in a (simplified) theoretical way and based on this can also provide some hints as to how to overcome some of the problems involved. To conclude the preliminary thoughts, without claiming completeness Table 1 surveys some of the studies dealing with problems related to the aggregation of time series.

2.2 The DGP

To keep things very simple we will investigate a simplest data generating process:

$$x_{1,s} = x_{1,s-1} + \varepsilon_{1,s} \tag{2.1}$$

$$x_{2,s} = x_{1,s-1} + \varepsilon_{2,s} \tag{2.2}$$

giving rise to the error correction representation

$$\triangle x_{1,s} = \varepsilon_{1,s} \tag{2.3}$$

$$\triangle x_{2,s} = -(x_{2,s-1} - x_{1,s-1}) + \varepsilon_{2,s} \tag{2.4}$$

with $\triangle_i^j = (1 - L^i)^j$ and L being the lag operator. The disturbances $\varepsilon_{1,s}$ and $\varepsilon_{2,s}$ are independent of each other and are assumed to follow a time invariant distribution with mean zero and variance σ_1^2 and σ_2^2 respectively. The difference $(x_{2,s-1} - x_{1,s-1})$ will be called error correction term (ec-term). Under these circumstances the processes are known to contain a unit root which becomes apparent from their stochastic

trend representations

$$x_{1,s} = \sum_{i=-\infty}^{s-1} \varepsilon_{1,i} + \varepsilon_{1,s} \tag{2.5}$$

$$x_{2,s} = \sum_{i=-\infty}^{s-1} \varepsilon_{1,i} + \varepsilon_{2,s} \tag{2.6}$$

forming the basis to re-write (2.4) as

$$\triangle x_{2,s} = -(\varepsilon_{2,s-1} - \varepsilon_{1,s-1}) + \varepsilon_{2,s}$$

In the terminology of Ericsson, Hendry & Mizon (1998) x_1 will be called weakly exogenous because while x_2 will be referred to as the endogenous variable in the bi-variate system.

2.3 Aggregation Strategies

We will consider two aggregation strategies, averaging and skip sampling. The following scheme makes clear what they mean.

DGP	$\underbrace{x_{i,s-8}\ x_{i,s-7}}$	$\underbrace{x_{i,s-6}\ x_{i,s-5}}$	$\underbrace{x_{i,s-4}\ x_{i,s-3}}$	$\underbrace{x_{i,s-2}\ x_{i,s-1}}$	$\underbrace{x_{i,s}\ x_{i,s+1}}$
averaging	$\underbrace{y_{i,r-4}}$	$y_{i,r-3}$	$\underbrace{y_{i,r-2}}$	$y_{i,r-1}$	$\underbrace{y_{i,r}}$
skip sampling	$z_{i,t-2}$		$z_{i,t-1}$		$z_{i,t}$

We define $y_{i,r} = \frac{1}{2}(x_{i,s} + x_{i,s+1})$ and $z_{i,t} = y_{i,r}$, $z_{i,t\pm1} = y_{i,r\pm2}$ which restricts the current analysis to the case of averaging over two neighbouring periods and by skip sampling to collecting every second data point ignoring the intermediate one.[3] Thus, skip sampling is the equivalent to what Granger & Siklos (1995) refer to as systematic sampling and averaging is used as an other expression for temporal aggregation. This choice has been made because the last term seems to be too general since skip sampling also can be regarded a particular version of temporal aggregation where unequal weights are attached to different observations. Likewise, systematic sampling appears less appealing than skip sampling which provides a more intuitive access to what is happening with the data.

[3]It is easy to see though that this procedure will generalize to averages of even numbers of periods at least.

2.3.1 Averaging

Now we first consider pure averaging and look at $y_{i,r}$ to start with.

$$\begin{aligned}
y_{1,r} &= \frac{1}{2}(x_{1,s} + x_{1,s+1}) \\
&= \frac{1}{2}\left(\sum_{i=0}^{s} \varepsilon_{1,i} + \sum_{i=0}^{s} \varepsilon_{1,i} + \varepsilon_{1,s+1}\right) \\
&= x_{1,s} + \frac{1}{2}\varepsilon_{1,s+1} \\
y_{2,r} &= \frac{1}{2}\left(\sum_{i=0}^{s-1} \varepsilon_{1,i} + \varepsilon_{2,s} + \sum_{i=0}^{s-1} \varepsilon_{1,i} + \varepsilon_{1,s} + \varepsilon_{2,s+1}\right) \\
&= x_{1,s-1} + \frac{1}{2}\left(\varepsilon_{1,s} + \varepsilon_{2,s} + \varepsilon_{2,s+1}\right)
\end{aligned}$$

$$\begin{aligned}
\triangle y_{1,r} &= x_{1,s} - x_{1,s-2} + \frac{1}{2}(\varepsilon_{1,s+1} - \varepsilon_{1,s-1}) \\
\triangle y_{2,r} &= x_{1,s-1} + \frac{1}{2}\left(\varepsilon_{1,s} + \varepsilon_{2,s} + \varepsilon_{2,s+1}\right) \\
&\quad - \left(x_{1,s-3} + \frac{1}{2}\left(\varepsilon_{2,s-2} + \varepsilon_{1,s-2} + \varepsilon_{2,s-1}\right)\right)
\end{aligned}$$

These expressions simplify to

$$\begin{aligned}
\triangle y_{1,r} &= \triangle_2 x_{1,s} + \varepsilon^*_{1,s} \\
\triangle y_{2,r} &= \triangle_2 x_{1,s-1} + \varepsilon^*_{2,s}
\end{aligned}$$

with $\varepsilon^*_{1,s} = \frac{1}{2}\left(1 - L^2\right)\varepsilon_{1,s+1}$ and $\varepsilon^*_{2,s} = \frac{1}{2}\left(1 - L^2\right)\left(\varepsilon_{1,s} + \varepsilon_{2,s} + \varepsilon_{2,s+1}\right)$. Re-arranging terms finally leads to an error correction representation of the averaged process

$$\triangle y_{1,r} = u_{1,r} \tag{2.7}$$

$$\triangle y_{2,r} = -\left(y_{2,r-1} - y_{1,r-1}\right) + u_{2,r} \tag{2.8}$$

where $u_{1,r}$ and $u_{2,r}$ are given by

$$u_{1,r} = \varepsilon_{1,r} + \frac{1}{2}\left(1 + L\right)\varepsilon_{1,s} \tag{2.9}$$

$$u_{2,r} = \varepsilon_{2,r} + \frac{1}{2}\left(1 + L\right)\varepsilon_{1,s} \tag{2.10}$$

and $\varepsilon_{i,r} = \frac{1}{2}\left(1 + L\right)\varepsilon_{i,s+1}$.

Equation (2.7) and (2.8) reveal three effects of the averaging procedure. At first, the disturbances of equation (2.7) feature moving average (MA) properties because some of the components of $u_{1,r}$ are also present in $u_{1,r-1}$.

- **Moving Average effect**

$$\begin{aligned} E(u_{1,r}u_{1,r-1}) &= E\left(\frac{1}{2}\varepsilon_{1,s-1}\right)^2 \\ &= \frac{1}{4}\sigma_1^2 \end{aligned}$$

In contrast to the first equation, no MA effect is introduced to the second one. The result also suggests that the order of the moving average process introduced is just one. Disturbances that are further apart than one period will not be correlated. Two more remarks are in order. The standard results for aggregated stable *VARMA(p,q)* DGP do not distinguish between the individual processes with respect to the MA effect.[4] In other words, in those cases all marginal processes feature MA components after aggregation. This is not the case here, however. The reason for that lies in the particular way in which lagged values of the observations enter the DGP. Second, also owed to the cointegration representation it is not possible to express the error term $u_{1,r}$ as a sum of two subsequent and independent error components, thus we can only talk of a MA effect rather than giving a precise MA representation of the error term.

The next consequence of aggregation has to be mentioned only briefly, because it is also present in the *VARMA* case and therefore does not come as a surprise. This effect is the correlation between the error terms of the individual processes. Some more detailed analysis thereof is available in Hallerbach (2000). While not specifying a data generating process and not considering cointegrated variables, he investigates averages of stochastic variables in general. One should recall that we assumed the innovations of the DGP do be independent of each other. This property obviously does not carry over to the averaged series while the auto-regressive structure of the process itself is not altered.

Due to the third effect the disturbances in (2.8) will be correlated with the explanatory variables, that is with the equilibrium error $ec_r = (y_{2,r-1} - y_{1,r-1})$. As a consequence ordinary least square estimation of that equation would yield biased

[4] see e.g. Lütkepohl (1993) p. 233.

estimates and standard parameter tests become invalid. It also calls into question the validity of general cointegration test e.g. system cointegration tests. Luckily, the small simulation study below indicates that standard system cointegration tests[5] do not seem to be affected at least as long as the DGP is as simple as in (2.1) and (2.2). The expected additional covariances introduced by the averaging procedure can be calculated as follows.

- **Covariance Regressor, disturbance**

$$\begin{aligned} E(ec_r u_{2,r}) &= E\left[\left(-\frac{1}{2}(1+L)\,\varepsilon_{1,s-1}+\frac{1}{2}(1+L)\,\varepsilon_{2,s-1}\right)\right. \\ &\quad \left.\times\left(\frac{1}{2}(1+L)\,\varepsilon_{1,s}+\frac{1}{2}(1+L)\,\varepsilon_{2,s+1}\right)\right] \\ &= E\left(-\frac{1}{2}\varepsilon_{1,s-1}\frac{1}{2}\varepsilon_{1,s-1}\right) \\ &= -\frac{1}{4}\sigma_1^2 \end{aligned}$$

- **Covariance of the disturbances**

$$\begin{aligned} E(u_{1,r}u_{2,r}) &= E\left[\left(\frac{1}{2}\varepsilon_{1,s-1}+\varepsilon_{1,s}+\frac{1}{2}\varepsilon_{2,s-1}\right)\right. \\ &\quad \left.\times\left(\frac{1}{2}(1+L)\,\varepsilon_{1,s}+\frac{1}{2}(1+L)\,\varepsilon_{2,s+1}\right)\right] \\ &= E\left(\frac{1}{2}\varepsilon_{1,s-1}+\frac{1}{2}\varepsilon_{1,s}\right)^2 \\ &= \frac{1}{2}\sigma_1^2 \end{aligned}$$

Note that in regression analysis the correlation between disturbance and regressor in the equation for $y_{2,t}$ can neither be accounted for by lagged values of the dependent variable nor by introducing MA terms. This is because the latter would not share any component with $u_{2,r}$. At the end of the paper a few remarks will be made with respect to a more general approach towards inclusion of lagged variables.

[5] see e.g. Johansen (1995)

2.3.2 Skip Sampling

As mentioned before, the phenomenon of the moving average properties of the series derived by aggregation have long been recognized. One seemingly obvious and popular solution to this problem is skip sampling.[6] That is, data which is known to have been averaged is selected such that some intermediate observations are thrown away and only values picked in this way are considered. For example data on interest rates which can be observed on a daily basis are summarized to monthly data and for econometric modelling on a quarterly frequency the average of the mid-quarter month is picked to represent the whole quarter. To analyze the effects we will proceed in two steps. First we calculate the results of pure skip sampling and second these will be combined with the previous results.

The following rules apply

$$\begin{aligned} y_{i,r} &= x_{i,s}, \quad s = r, \\ z_{i,t\pm1} &= y_{i,s\pm2} \end{aligned}$$

Straightforward re-arrangement gives

$$\begin{aligned} \triangle z_{1,t} &= e_{1,t} && (2.11) \\ \triangle z_{2,t} &= -(\varepsilon_{2,s-2} - \varepsilon_{1,s-2}) + (\varepsilon_{1,s-1} + \varepsilon_{2,s}) \\ &= -(z_{2,t-1} - z_{1,t-1}) + e_{2,t} && (2.12) \end{aligned}$$

where

$$\begin{aligned} e_{1,t} &= \varepsilon_{1,s-1} + \varepsilon_{1,s} && (2.13) \\ e_{2,t} &= \varepsilon_{1,s-1} + \varepsilon_{2,s} && (2.14) \end{aligned}$$

from which we can conclude that the new error terms will indeed not be correlated over time, that means there is no moving average effect introduced as long as $E(\varepsilon_{1,s-2}\varepsilon_{1,s}) = E(\varepsilon_{2,s-2}\varepsilon_{2,s-1}) = 0$.

Moreover, there is also no correlation between the errors of equation (2.12) and the explanatory variables. On the other hand the disturbances of both equations

[6] Wolters, Teräsvirta & Lütkepohl (1998) do so for example.

will be correlated, the covariance between them being σ_1^2. Nevertheless, the error correction representation still exists, so skip sampling the data does not have an effect on that.

2.3.3 Combining Skip Sampling and Averaging

The last computational effort will be devoted to combine the averaging and the skip sampling procedure. We will now skip sample the averaged process. We already know that the error correction representation of the skip sampled process is the same as for the original process which is in our case the averaged process. We therefore simply replace the components of the screened representation by the counterparts from the averaged results. As a consequence the properties of the error terms will be different. We re-state equations (2.11) and (2.12) noting that we now have $y_{i,r} = \frac{1}{2}(x_{i,s} + x_{i,s+1})$, $z_{i,t} = y_{i,r}$ and $z_{i,t\pm1} = y_{i,r\pm2}$ and so forth, we find from equations (2.7) and (2.8):

$$\triangle z_{1,t} = e^*_{1,t} \tag{2.15}$$

$$\triangle z_{2,t} = -(z_{2,r-1} - z_{1,r-1}) + e^*_{2,t} \tag{2.16}$$

As a consequence of (2.13) and (2.14) we are able to dissolve $e_{1,t}$ and $e_{2,t}$ as

$$e^*_{1,t} = u_{1,r-1} + u_{1,r} \tag{2.17}$$

$$e^*_{2,t} = u_{1,r-1} + u_{2,r} \tag{2.18}$$

Likewise the term $-(z_{2,t-1} - z_{1,t-1})$ can be found to be

$$-(z_{2,t-1} - z_{1,t-1}) = -u_{2,r-2} + u_{1,r-2} \tag{2.19}$$

Thus, expressions for $u_{i,r}$ will provide the solution. We have

$$u_{1,r} = \frac{1}{2}(\varepsilon_{1,s-1} + \varepsilon_{1,s+1}) + \varepsilon_{1,s} \tag{2.20}$$

$$u_{2,r} = \frac{1}{2}(\varepsilon_{1,s-1} + \varepsilon_{1,s}) + \frac{1}{2}(\varepsilon_{2,s} + \varepsilon_{2,s+1}) \tag{2.21}$$

Carefully respecting the step length of the skip sampling procedure further reveals that

$$u_{1,r-1} = \frac{1}{2}(\varepsilon_{1,s-3} + \varepsilon_{1,s-1}) + \varepsilon_{1,s-2}$$

$$
\begin{aligned}
u_{1,r-2} &= \frac{1}{2}\left(\varepsilon_{1,s-5}+\varepsilon_{1,s-3}\right)+\varepsilon_{1,s-4} \\
u_{2,r-2} &= \frac{1}{2}\left(\varepsilon_{1,s-5}+\varepsilon_{1,s-4}\right)+\frac{1}{2}\left(\varepsilon_{2,s-4}+\varepsilon_{2,s-3}\right)
\end{aligned}
$$

We are now in a position to calculate the items of interest. First re-consider the moving average problem. After dismissing the next observation, that is choosing $z_{t+1} = y_{r+2}$, the new error $e^*_{1,t+1}$ is given by $e^*_{1,t+1} = u_{1,r+1} + u_{1,r+2}$ which shares the (s+1)st innovation $\varepsilon_{1,s+1}$ with $e^*_{1,t}$. That's why it does not suffice to drop only one observation and in fact, there is no skip length that could guarantee removal of the MA pattern at all. The reason for this can be found in the following decomposition of a random walk where we intend to pick every ath observation only ($a > 1$).

$$
\begin{aligned}
x_t &= \sum_{i=-\infty}^{t-1}\varepsilon_i+\varepsilon_t \\
&= \sum_{i=-\infty}^{t-a}\varepsilon_i+\sum_{i=t-a+1}^{t-1}\varepsilon_i+\varepsilon_t \\
&= x_{t-a}+e^*_t
\end{aligned}
\tag{2.22}
$$

It tells that regardless of the number of observations being not considered e.g. by leaving them out, the ignored effects will turn up in the error term. These are subject to an MA(1) effect when previously averaged observation enter (2.22), however. Thus, due to ε_{t-a} and ε_t the errors e^*_{t-a} and e^*_{t+a} respectively will be correlated with e^*_t.

We conclude that skip sampling does not remove the moving average effects introduced by the averaging procedure while maintaining the cointegration structure of the model. Secondly, the covariance between the innovations in (2.15) and (2.16) is given by

$$
E(e^*_{1,t}e^*_{2,t}) = \frac{11}{4}\sigma_1^2 \tag{2.23}
$$

where we made use of the fact that

$$
\begin{aligned}
e^*_{1,t} &= \frac{1}{2}\varepsilon_{1,s-3}+\varepsilon_{1,s-2}+\varepsilon_{1,s-1}+\varepsilon_{1,s}+\frac{1}{2}\varepsilon_{1,s+1} \\
e^*_{2,t} &= \frac{1}{2}\varepsilon_{1,s-3}+\varepsilon_{1,s-2}+\varepsilon_{1,s-1}+\frac{1}{2}\varepsilon_{1,s}+\frac{1}{2}\varepsilon_{2,s}+\frac{1}{2}\varepsilon_{2,s+1}
\end{aligned}
$$

Note therefore that the data screen does not improve the situation with respect to the artificially introduced covariance. Finally, the covariance between regressor and disturbance in (2.16) is investigated:

$$E\left((z_{2,t-1} - z_{1,t-1})e^*_{2,t}\right) = -\frac{1}{4}\sigma_1^2 \tag{2.24}$$

Thus the explanatory variables remain correlated with the error terms in that equation generally jeopardising standard inference in an OLS framework.

Table 2 states the results for the three processes considered. These are the original process, the purely averaged, the purely screened and the averaged and screened process. For convenience we will label the number of observations entering the average m, and the observations skipped to obtain the screened process $a-1$. So, we will find the averaged and screened process for e.g. $m = a = 1$ to be the original sample.

2.4 A small simulation study

Based on the considerations in the previous sections we will simulate some models and check the performance of testing procedures. In line with model (2.4) we formulate its empirical counterpart as

$$\triangle z_{2,t} = \alpha(z_{2,t-1} - z_{1,t-1}) + e^*_{2,t} \tag{2.25}$$

where the $z_{i,t}$ had previously been subject to the data transformation procedures described by a and m. That means we extend the averaging and skip sampling technology to encompass cases which are common in praxis. As to the equation for the $z_{1,t}$ we do not perform estimation, but we make use of them to calculate the correlations between the errors of both processes. That is calculating the correlation between $\triangle z_{1,t}$ and $(\triangle z_{2,t} - \hat{\alpha}(z_{2,t-1} - z_{1,t-1}))$, where $\hat{\alpha}$ is the OLS estimate of α in equation (2.25). Likewise, the quantities $(z_{2,t-1} - z_{1,t-1})$ and $(\triangle z_{2,t} - \alpha(z_{2,t-1} - z_{1,t-1}))$ are used to calculate the correlations between the regressor and the error term in (2.25). It has been mentioned before that the latter effect will bias the OLS estimates of α, therefore we will check this bias too. We formulate two hypotheses: $H_0^1 : \alpha = 0$

Table 2: Covariances and correlation coefficients

	error,regressor		between errors	
	Cov.	Corr.	Cov.	Corr.
		formulas		
original data (m=a=1)	0	0	0	0
(I) averaging (m=2, a=1)	$-\frac{1}{4}\sigma_1^2$	$\frac{-\frac{1}{4}\sigma_1^2}{\frac{1}{2}\left(\sigma_1^2+\sigma_2^2\right)}$	$\frac{1}{2}\sigma_1^2$	$\frac{\frac{1}{2}\sigma_1^2}{\sqrt{\frac{3}{4}\sigma_1^2}\sqrt{\frac{1}{2}(\sigma_1^2+\sigma_2^2)}}$
(II) skip sampling (m=1, a=2)	0	0	σ_1^2	$\frac{\sigma_1^2}{\sqrt{2\sigma_1^2}\sqrt{\sigma_1^2+\sigma_2^2}}$
(I) + (II) (m=2, a=2)	$-\frac{1}{4}\sigma_1^2$	$\frac{-\frac{1}{4}\sigma_1^2}{\sqrt{\frac{1}{2}(\sigma_1^2+\sigma_2^2)}\sqrt{\frac{5}{2}\sigma_1^2+\frac{1}{2}\sigma_2^2}}$	$\frac{11}{4}\sigma_1^2$	$\frac{\frac{11}{4}\sigma_1^2}{\sqrt{\frac{7}{2}\sigma_1^2}\sqrt{\frac{5}{2}\sigma_1^2+\frac{1}{2}\sigma_2^2}}$
		Evaluation at $\sigma_1^2=\sigma_2^2=1$		
original data (m=a=1)	0	0	0	0
(I) averaging (m=2, a=1)	$-\frac{1}{4}$	$-\frac{1}{4}$	$\frac{1}{2}$	0.577
(II) skip sampling (m=1, a=2)	0	0	1	.5
(I) + (II) (m=2, a=2)	$-\frac{1}{4}$	$-.144$	$\frac{11}{4}$	0.849
		Simulation results*, no estimation		
original data (m=a=1)		.001		.001
(I) averaging (m=2, a=1)		-.250		.611
(II) skip sampling (m=1, a=2)		.002		.500
(I) + (II) (m=2, a=2)		-.145		.848
		Simulation results*, α, $e_{2,t}$ estimated		
original data (m=a=1)				0
(I) averaging (m=2, a=1)				0.579
(II) skip sampling (m=1, a=2)				0.500
(I) + (II) (m=2, a=2)				0.837

*Simulation with $R = 1000$ draws and $T = 1000$ observations.

being equivalent to hypothesizing no cointegration and second $H_0^2 : \alpha = -1$ to actually asses the deviation of the estimate from its true value. Finally, stacking $z_{1,t}$ and $z_{2,t}$ into an vector $z_t = (z_{1,t}\ z_{2,t})'$, we can write the model in the vector error correction form[7]:

$$\triangle z_t = \alpha\beta' z_{t-1} + \varepsilon_t \tag{2.26}$$

allowing us to perform system cointegration tests. We chose the Johansen[8] tests because in this very basic framework there is no need for more sophisticated modelling and these tests are the most commonly used anyway.

The following values of a and m are chosen. Apart from the 1, 2 representing the situations we have theoretical results for, m will take on the values 7, 14, 28, 84 to approximate weekly, fortnightly, monthly and quarterly averages of daily data respectively, a also becomes 3, 4, 16, 28 and 84 to allow picking monthly and quarterly representatives of the averaged observations. The combination $m = 28,\ a = 3$ will approximate the situation where we pick the end of month value within a quarter of monthly averages, for example.

The experiment is run a $R = 1000$ times with $T = 100$ observations each[9], and the starting seed for the GAUSS NT 3.2 version random normal number generator is set to 8091976. In addition, 1000 draws of the random numbers have been reserved as pre-sample values and therefore do not enter the simulation. The error terms are (pseudo-) normally distributed with mean zero and unit variance, yet independent of each other.

Although we do not have theoretical results for the cases involving higher order m and a aggregation, the general suspicion is that the properties derived for the basic situations would carry over to those. And indeed, that seems to be the case. Two important questions have to be addressed. First the cointegration property and second the estimate of the adjsutment parameter. Unfortunately, we cannot easily check both, estimates of the cointegration and adjustment parameter.

[7] The true parameters of α and β are $(0\ 1)'$ and $(1\ -1)$ respectively.

[8] see Johansen (1988)

[9] The theoretical results for the correlations are simulated with $T = 1000$ observations too.

Table A.2 can be interpreted as the results of a test for cointegration. If the adjustment parameter really was zero, this would imply no cointegration. Rejecting the hypothesis $\alpha = 0$ therefore provides evidence for both processes being cointegrated. That's why ideally all entries should be close to one. It turns out that the correct decision becomes the less likely the more involved the averaging procedure becomes. For the combination approximating the quarterly representative of monthly averages ($m = 28,\ \ a = 3$), no cointegration is found in about one third of all throws. This is the more surprising since the bias introduced in the OLS procedure is a negative one, therefore, the estimates of α are skewed downwards from minus one and thus further away from zero. The same effect becomes apparent from table A.3 where the bias is investigated more thoroughly. In contrast to the previous aspect, where the test seemed quite robust against modest manipulations of the data, even very limited aggregation operations heavily reduce the likelihood to obtain correct estimates of the adjustment parameter. Averaging and skip sampling over just two periods each e.g., exaggerates the nominal size by more than .25. This effect seems to be much stronger the more consecutive observations are averaged while skip sampling relieves some of the distortions. The reason for this behaviour lies in the fact that the correlation between the error term and the regressor of the second variable becomes smaller the larger the steps of the screen.

With respect to the correlation between the residuals of both equations, table A.1 indicates that averaging and skip sampling both seem to feedback to each other. Therefore, the largest correlation coefficients are obtained for high orders of a and m.

Finally, having mentioned that cointegration may not be found on grounds of single equation OLS regression, the system cointegration test seem to solve that problem as can be seen from table A.4. We find that the selection of the cointegrating rank succeeds in all cases at all reasonable significance levels. Thus, aggregation alone cannot be blamed for not finding cointegration between inflation rates.

2.5 More Questions and Few More Answers

In the tradition of the previous studies this paper investigated those situations where all series involved are aggregated by either averaging or skip sampling. The combination of both has not yet been dealt with in detail. This however, is not the whole universe of possible combinations which may occur in praxis. One such typical constellation is the joint empirical analysis of skip sampled and averaged variables when for example the demand for money is modelled as a function of income and interest rates.[10] Therefore, the consequences of these possibilities deserve a separate treatment. Of course, the approach presented above is capable of also deriving the theoretical results for these situations. The only difference is that it is now necessary to distinguish between the weakly exogenous and the endogenous variable. To save space, the straightforward calculations are omitted and the results are presented in Table 3. It turns out that the least problematic case is the one in which the weakly exogenous process is skip sampled while the other may be averaged. In fact, the effects are mainly identical independent of whether or not both processes are skip sampled or if the weakly exogenous process is skip sampled and the other is averaged.

As shown in Tables 2 and A.3 the most troublesome problem seems to be the correlation between the error correction term and the regression residual in the equation for the endogenous variable. A natural candidate for solving this problem is lag augmentation of the regression equations. Looking at the Tables 3,4,5 indicates that including the one time lagged dependent weakly exogenous variable in the equation for the other variable suffices to eliminate the correlation between the error correction term and the residuals. Scrutinising this recipe more thoroughly reveals that this incurs costs in its own, though. In all cases the residual variance will increase and in some instances the MA effect will become more pronounced. In only one of the constellations the correlation between the residuals will be reduced

[10] See for example Breitung & Swanson (1998) who point out that flow variables like income can be viewed as averaged observations while skip sampling is more likely applied to stock variable like money.

Table 3: Different Aggregation Strategies and the Implications for the DGP Representation and Error Correlations

$$\begin{aligned} \Delta y_{1,t} &= e_{1,t} \\ \Delta y_{2,t} &= \alpha(y_{1,t-1} - y_{2,t-1}) + e_{2,t} \end{aligned}$$

Effect on	process: weakly exogenous, skip sampled	process: weakly exogenous, averaged	process: endogenous
Representation	preserved	preserved	skip sampling
correlations			
ec-term, e_{2t}	0	x	
$e_{1,t}, e_{2,t}$	x	x	
$e_{i,t}, e_{i,t\pm1}$	0	x/0	
Representation	preserved	preserved	averaging
correlations			
ec-term, e_{2t}	0	x	
$e_{1,t}, e_{2,t}$	x	x	
$e_{i,t}, e_{i,t\pm1}$	0	x	

* The x indicate non-zero effects, x/0 stands for a non-zero correlation in the weakly exogenous process only. Note: if no aggregation would precede the empirical analysis all correlations would be zero.
ec-term, $e_{i,t}$ stand for their respective counterparts in the representation of the data generating process.

Table 4: The Consequences of Lag Augmentation

$$\begin{aligned} \triangle y_{1,t} &= e_{1,t} \\ \triangle y_{2,t} &= \alpha(y_{1,t-1} - y_{2,t-1}) + \gamma \triangle y_{1,t-1} + e_{2,t} \end{aligned}$$

Effect on	process: weakly exogenous, skip sampled	process: weakly exogenous, averaged	process: endogenous
Representation	preserved	preserved	
correlations			
ec-term, e_{2t}	0	0	skip sampling
$e_{1,t}, e_{2,t}$	0	x	
$e_{i,t}, e_{i,t\pm1}$	0	**x**/0	
Representation	preserved	preserved	
correlations			
ec-term, e_{2t}	0	0	averaging
$e_{1,t}, e_{2,t}$	x	**x**	
$e_{i,t}, e_{i,t\pm1}$	0	x	

* The x indicate non-zero effects, x/0 stands for a non-zero correlation in the weakly exogenous process only. Bold face marks stronger correlation if compared to Table 3.

ec-term, $e_{i,t}$ stand for their respective counterparts in the representation of the data generating process.

to zero if the appropriate lagged variable was included in the regression.

Finally, if the regression equation for the series obtained by applying both averaging and skip sampling is augmented by a lagged variable, the correlation between the error correction term and the innovations will disappear but the newly introduced regressor will be correlated with both, the error term and the error correction term. Thus, as before one problem is solved at the expense of introducing another one.

Table 5: Skip Sampling the Averaged Process and Lag Augmentation

$$\begin{aligned} \triangle y_{1,t} &= e_{1,t} \\ \triangle y_{2,t} &= \alpha(y_{1,t-1} - y_{2,t-1}) + \gamma \triangle y_{1,t-1} + e_{2,t} \end{aligned}$$

Effect on	m = 2, a = 2	
	$\gamma = 0$	$\gamma \neq 0$
Representation	preserved	preserved
correlations		
ec-term, e_{2t}	x	0
$e_{1,t}, e_{2,t}$	x	x
$e_{i,t}, e_{i,t\pm1}$	x	x
$\triangle y_{1,t-1}, e_{2,t}$		x
$\triangle y_{1,t-1}, (y_{1,t-1} - y_{2,t-1})$		x

* The x indicate non-zero effects, x/0 stands for a non-zero correlation in the weakly exogenous process only. Bold face marks stronger correlation if compared to Table 3.

ec-term, $e_{i,t}$ stand for their respective counterparts in the representation of the data generating process.

3 Conclusions

We investigated a small, specific, cointegrated process which has been subjected to standard aggregation procedures commonly used in preparation for econometric analysis of macro-economic data. In contrast to previous studies we focused on the cointegrating feature of the data generating process and the implications for the relation between the variables rather than for the isolated processes alone.

It turned out that averaging does not always cause moving average effects. In our example this was the case for the independent random walk only, while the series being ruled by the error correction term did not suffer from this problem. In addition, in contrast to the *VARMA(p,q)* case, the problem cannot be solved by augmenting the cointegrated process by a MA term.

Regardless of the aggregation method (averaging or skip sampling), the innovations of the aggregated processes will be correlated even if those of the underlying data generating processes are not. This is useful knowledge for e.g. impulse-response analysis of vector-autoregressive processes (VAR), in particular if the error covariances are given an economic interpretation like in the structural VAR analysis. In such cases spurious correlation might render the identification of the structural shocks arbitrary. In the more general perspective, the conclusions drawn with respect to instantaneous causality have to be met with caution.

Second, if the process is the result of averaging, the innovations of the data series being ruled by the exogenous stochastic trend (here $z_{2,t}$) become correlated with the error correction term entering the same equation. Standard OLS analysis therefore yields biased estimates of the adjustment coefficient. That's why cointegration tests in the single equation framework frequently fail, though less so than finding the true parameter itself. It does not occur if the data is screened only and it is relieved if the data is screened once it has been averaged before.

Finally, standard system cointegration analysis opens a way to cope with at least some of the problems. They seem to be reliable when it comes to determination of the cointegrating rank. When including lags of the weakly exogenous variable as

regressors to the equation for the endogenous variable, standard OLS inference will improve with respect to the ec-term but only at the expense of introducing other problems.

So far, the outcome cannot be generalized to all cointegrated processes and richer lag structures, yet one should be prepared to face similar problems. Therefore, a natural extension of the current agenda is to generalize the investigation to encompass larger cointegration parameter spaces. In addition, general theoretic results concerning the error correlation are desirable to provide some form of correction for that phenomenon.

References

Breitung, J. & Swanson, N. R. (1998). Temporal Aggregation and Causality in Multiple Time Series Models, *Discussion Paper 27*, Sonderforschungsbereich 373, Humboldt-Universität zu Berlin.

Ericsson, N. R., Hendry, D. F. & Mizon, G. (1998). Exogeneity, Cointegration, and Economic Policy Analysis, *Journal of Business and Economic Statistics* **16**(4): 370 – 87.

Gottschalk, J. (1999). A Cointegration Analysis of a Money Demand System in Europe, *Kiel Working Paper 902*, The Kiel Institute of World Economics.

Granger, C. W. & Siklos, P. L. (1995). Systematic sampling, temporal aggregation, seasonaladjustment, and cointegration Theory and evidence, *Journal of Econometrics* **66**: 357 – 69.

Hallerbach, W. G. (2000). Cross- and auto-correlation effects arising from averaging, *Discussion Paper TI 2000-064/2*, Tinbergen Institute.

Hamilton, J. D. (1994). *Time Series Analysis*, 1st edn, Princeton University Press, Princeton, New Jersey USA.

Johansen, S. (1988). Statistical Analysis of Cointegration Vectors, *Journal of Economic Dynamics* **12**: 231 – 54.

Johansen, S. (1995). *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*, 1st edn, Oxford University Press, Oxford.

Lütkepohl, H. (1993). *Introduction to Multiple Time Series Analysis*, 2nd edn, Springer-Verlag, Berlin.

Mamingi, N. (1996). Aggregation over time, error correction models and granger causality: A Monte Carlo investigation, *Economics Letters* **52**: 7 – 14.

Müller, C. & Hahn, E. (2001). Money Demand in Europe: Evidence from the Past, *Kredit und Kapital* **34**(1): 48 – 75.

Tiao, G. (1972). Asymptotic behaviour of temporal aggregates of time series, *Biometrica* **59**(3): 525 – 31.

Wolters, J., Teräsvirta, T. & Lütkepohl, H. (1998). Modelling the Demand for M3 in the Unified Germany, *Review of Economics and Statistics* **80**(3): 399–409.

Working, H. (1960). Note on the correlation of first differences od averages in random chain, *Econometrica* **28**(4): 916 – 918.

A Appendix

Table A.1: Empirical correlations between the error terms

T=100		averaging parameter (m)					
R=1000		1	2	7	14	28	84
	1	.001	.570	.894	.905	.903	.900
	2	.484	.831	.955	.958	-	-
skip	3	.662	-	-	-	.972	-
parameter (a)	4	.743	-	.976	-	-	-
	16	.932	-	-	-	-	.991
	28	.959	-	-	-	-	-
	84	.983	-	-	-	-	-

The α have been estimated by OLS previously.

Table A.2: Cointegration test: Empirical rejection frequencies of $H_0^1 : \alpha = 0$

T=100		averaging parameter (m)					
R=1000		1	2	7	14	28	84
	1	1	1	1	1	.999	.999
	2	1	1	.981	.942	-	-
skip	3	1	-	-	-	.694	-
parameter (a)	4	1	-	.774	-	-	-
	16	.934	-	-	-	-	.171
	28	.778	-	-	-	-	-
	84	.324	-	-	-	-	-

The nominal significance level is 0.05.

Table A.3: OLS bias analysis: Empirical rejection frequencies of $H_0^2 : \alpha = -1$

T=100		averaging parameter (m)					
R=1000		1	2	7	14	28	84
	1	.049	.758	.991	.995	.995	.996
	2	.074	.319	.722	.746	-	-
skip	3	.049	-	-	-	.575	-
parameter (a)	4	.059	-	.353	-	-	-
	16	.055	-	-	-	-	.165
	28	.049	-	-	-	-	-
	84	.052	-	-	-	-	-

The nominal significance level is 0.05.

Table A.4: System Cointegration Test: Empirical frequencies of correct rank decisions*

T=100		averaging parameter (m)					
R=1000		1	2	7	14	28	84
	1	.945	.943	.960	.956	.959	.962
	2	.936	.951	.954	.963	-	-
skip	3	.945	-	-	-	.951	-
parameter (a)	4	.947	-	.957	-	-	-
	16	.944	-	-	-	-	.963
	28	.959	-	-	-	-	-
	84	.950	-	-	-	-	-

*Johansen rank test, sequential test procedure.

The nominal significance level is 0.05.