ECB EZB EKT BCE EKP

# INTERPOLATION AND BACKDATING WITH A LARGE INFORMATION SET

**BY ELENA ANGELINI, JÉRÔME HENRY AND MASSIMILIANO MARCELLINO**
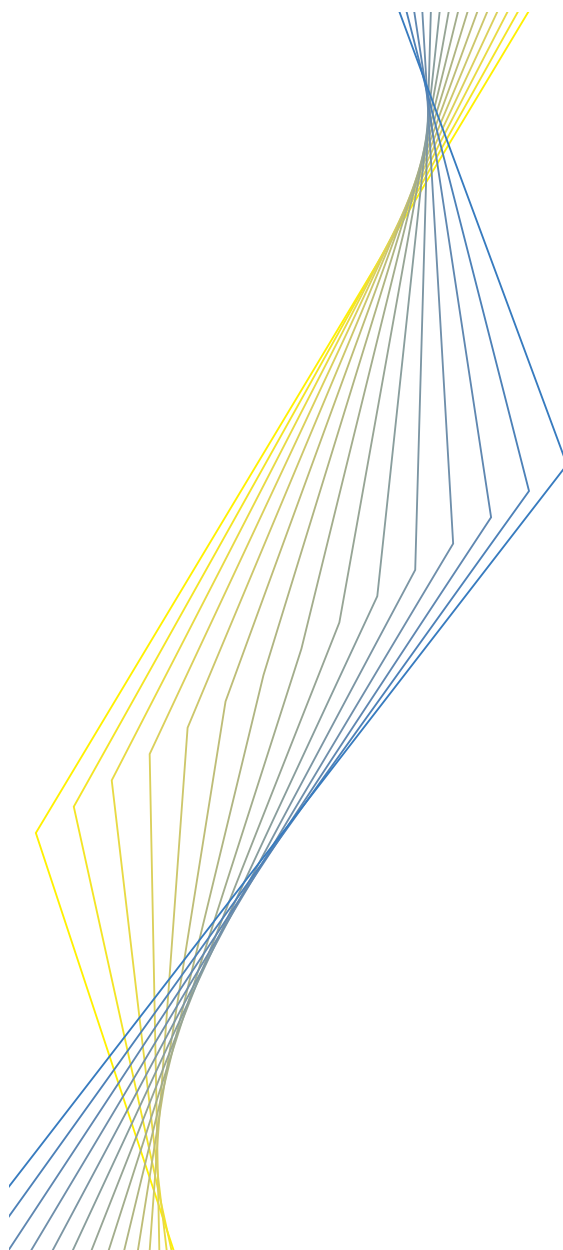
# WORKING PAPER NO. 252

# INTERPOLATION AND BACKDATING WITH A LARGE INFORMATION SET[1]

# BY ELENA ANGELINI[2], JÉRÔME HENRY[2] AND MASSIMILIANO MARCELLINO[3]

# AUGUST 2003

# Contents

**Abstract**

Existing methods for data interpolation or backdating are either univariate or based on a very limited number of series, due to data and computing constraints that were binding until the recent past. Nowadays large datasets are readily available, and models with hundreds of parameters are fastly estimated. We model these large datasets with a factor model, and develop an interpolation method that exploits the estimated factors as an efficient summary of all the available information. The method is compared with existing standard approaches from a theoretical point of view, by means of Monte Carlo simulations, and also when applied to actual macroeconomic series. The results indicate that our method is more robust to model misspecification, although traditional multivariate methods also work well while univariate approaches are systematically outperformed. When interpolated series are subsequently used in econometric analyses, biases can emerge, depending on the type of interpolation but again be reduced with multivariate approaches, including factor-based ones.

*Key words:* Interpolation, Factor Model, Kalman Filter, Spline

*JEL Classification:* C32, C43, C82.

**Non-technical summary**

A number of issues related to missing observations have received considerable attention in the econometric literature. Creating higher frequency data on the basis of available lower frequency series (e.g. deriving monthly values from quarterly values), interpolating missing observations or generating backdata when only recent observations are available (a task often called 'backcasting', as opposed to forecasting); all of these can be achieved using similar techniques. Existing methods however focus on one or on a small number of series, although a large amount of information is now readily available in the form of datasets with many variables for a considerable time span. The main statistical problem is then to find a proper way of exploiting these large datasets, in some sense by summarising them. This is now feasible, due to recent developments in the factor analysis literature.

In this paper we develop a factor-based approach to interpolation and estimation of missing observations that can exploit the information in very large datasets, and hence can be expected to perform better than existing limited information-based approaches. We then compare our method with existing alternatives from three perspectives: first, from a theoretical point of view; second, using artificial data, generated by means of Monte Carlo simulations, and, finally, in empirical applications - using actual macroeconomic variables of euro area countries.

The main results of the comparative analysis we conducted are the following.

First, the theoretical analysis indicates that large information sets are potentially useful, although the resulting estimators of missing observations are computationally not feasible unless some restrictions are imposed on the generating mechanism of the data. A reasonable assumption for economic time series is that they are driven by a very limited number of common forces, which in statistical terms implies that they can be represented by a factor model. The estimated factors represent an efficient summary of the information contained in the large dataset, and can be used to derive improved and feasible estimators of missing observations.

Second, Monte Carlo experiments in which deleted data from artificial series are re-estimated using several methods (Kalman filter and smoother, Spline, Chow and Lin, factor approach) suggest that when the series have themselves been generated by a factor process, the factor method tends to dominate all of the others, although the Chow and Lin method also performs well. Univariate techniques, on the contrary, yield poor results. When the data have been generated by univariate processes, as expected, univariate methods do the best job, in particular the Spline. The factor method however gives comparable results and, more importantly, it has to be borne in mind that real-life data is not very likely to follow such a simple data generating process.

Third, with actual time-series, namely quarterly GDP growth and inflation for 7 countries of the euro area, the results are similar to those obtained when the artificial data is generated by factors, i.e the multivariate methods clearly outperform the univariate ones. The Chow and Lin technique, in particular, delivers very good results overall, especially for inflation. One reason to explain this comparatively better performance, with respect to the factor method, is that the variables used in the Chow and Lin procedure were pre-selected according to the correlation with the series to be interpolated / backdated. Although this biases somewhat the experiment against the factor method, such an approach is however supposed to reflect practitioners' standard practice.

Finally, using both artificial and actual series, we evaluate the extent to which substituting the interpolated / backdated series to the original ones would affect both the estimated first order autocorrelation and a regression coefficient between two given series. The results are more favourable to the Chow and Lin technique, in particular for GDP growth. This presumably stresses again the importance of the pre-selection of most appropriate variables before running the interpolation procedure. An interesting general caveat emerging from this analysis is that biases resulting from the use of interpolated or backdated data can be sizeable, especially when there are a large number of missing observations.

# 1 Introduction

Issues of estimation of disaggregate data (e.g. monthly values from quarterly values), missing observations, and outliers received considerable attention in the literature. A first, simple, approach to recovering the disaggregated values is based on partial weighted averages of the aggregated ones, see e.g. Lisman and Sandee (1964). In a more sophisticated method, the disaggregated values are those which minimise a loss function under a compatibility constraint with aggregated data, see e.g. Boot *et al.* (1967), Cohen *et al.* (1971), Stram and Wei (1986). A further constraint can be added, the existence of a preliminary disaggregated series, so that the issue becomes how to best revise it in order for it to be compatible with the aggregated data, see e.g. Denton (1971), Chow and Lin (1971), Fernandez (1981), and Litterman (1983). The problem is somewhat simplified by assuming an ARIMA process at the disaggregate level, see e.g. Wei and Stram (1990) and Guerrero (1990). As far as the literature on missing observations and outliers is concerned, a selected list of references includes Harvey and Pierse (1984), Kohn and Ansley (1986), Nijman and Palm (1986), and Gomez and Maravall (1994).

All these methods, reviewed in Marcellino (1998), are univariate or only focus on a small number of series, while a large amount of information is now readily available in the form of datasets with many variables for a considerable time span. The main statistical problem is to find a proper representation for these large datasets, but recent developments in the factor analysis literature provide a solution. Standard factor models are not suited for applications with economic variables, since they require both the factors and the errors to be uncorrelated over time, and the errors to be orthogonal to each other. The latter hypothesis is relaxed in the static approximate factor model, see e.g. Chamberlain and Rothschild (1983), Connor and Korajczyk (1986, 1993). In the dynamic factor model the factors and the errors are also allowed to be correlated in time, see Stock and Watson (1998) and Forni, Hallin, Lippi and Reichlin (2000) for, respectively, a time domain and a frequency domain approach. The dynamic factor model has been shown to provide a proper representation for large dataset of macroeconomic variables, and in particular for forecasting, which can be considered as a problem of missing observations at the end of the series, see e.g. Stock

and Watson (1998) for the US, Marcellino, Stock and Watson (2001) and Angelini, Henry and Mestre (2001) for the Euro area, Artis, Banerjee and Marcellino (2001) for the UK. This suggests that similar methods could also be used to back-cast or backdate series for which information on the past is missing.

In this paper we develop a dynamic factor based approach to data interpolation and series backdating, compare it with existing methods from a theoretical point of view and by means of Monte Carlo simulations, and apply it to macroeconomic variables. More specifically, in section 2 we present the statistical framework. In section 3 we develop the factor based estimators, and compare them with competing methods from a theoretical point of view. In section 4 we evaluate the relative merits of the methods by means of simulation experiments. In section 5 we apply the methods to some macroeconomic variables. In section 6 we evaluate the consequences of using the interpolated / backdated data in subsequent analyses. Finally, in section 7 we summarize the main findings of the paper and conclude.

## 2   The Framework

We assume that the $n \times 1$ vector of weakly stationary time series $X_t$ admits the factor representation

$$\underset{n \times 1}{X_t} = \underset{n \times p}{\Lambda} \underset{p \times 1}{F_t} + \underset{n \times 1}{e_t} , \tag{1}$$

where $p$, the number of factors, is substantially smaller than $n$, namely, a few common forces drive the joint evolution of all the variables. Precise conditions on the factors, $F_t$, and the idiosyncratic errors, $e_t$, can be found in Stock and Watson (1998).

$y_t^o$ is a univariate series that can be also described by a factor structure

$$y_t^o = \beta' F_t + \varepsilon_t. \tag{2}$$

Yet, not all values of $y_t^o$ can be observed. In particular, observed values can be thought of as realizations of the process $y = \{y_\tau\}_{\tau=1}^\infty = \{\omega(L)y_{kt}^o\}_{t=1}^\infty$, where $\tau$ indicates the aggregate temporal frequency (e.g. quarters), $k$ the frequency of aggregation (e.g. 3 if $t$ is measured in months), $L$ is the lag operator, and $\omega(L) = \omega_0 + \omega_1 L + ... + \omega_{k-1} L^{k-1}$ characterizes the aggregation

scheme. For example, $\omega(L) = 1 + L + ... + L^{k-1}$ in the case of flow variables and $\omega(L) = 1$ for stock variables.

If we stack the observations on $X_t$, $y_t^o$, and $y_\tau$ in $\underset{nT \times 1}{\mathbf{X}}$, $\underset{T \times 1}{\mathbf{Y}^o}$ and $\underset{s \times 1}{\mathbf{Y}}$, where $s$ is the number of aggregate observations, and construct the aggregator matrix $\overline{\mathbf{W}}$ with

$$
\underset{(s+nT) \times (n+1)T}{\overline{\mathbf{W}}} = \begin{pmatrix} \underset{s \times T}{\mathbf{W}} & \underset{s \times nT}{\mathbf{0}} \\ \underset{nT \times T}{\mathbf{0}} & \underset{nT \times nT}{\mathbf{I}} \end{pmatrix},
$$

$$
\underset{s \times T}{\mathbf{W}} = \begin{pmatrix} \omega_0, \omega_1, ..., \omega_{k-1} & 0,0,...,0 & ... & 0,0,...,0 \\ 0,0,...,0 & \omega_0, \omega_1, ..., \omega_{k-1} & ... & 0,0,...,0 \\ ... & & & \\ 0,0,...,0 & 0,0,...,0 & & \omega_0, \omega_1, ..., \omega_{k-1} \end{pmatrix},
$$

then $\mathbf{Z} = \overline{\mathbf{W}}\mathbf{Z}^o$, where $\mathbf{Z}^o = (\mathbf{Y}^{o\prime} : \mathbf{X}')'$ and $\mathbf{Z} = (\mathbf{Y}' : \mathbf{X}')'$. The identity matrix in $\overline{\mathbf{W}}$ can be substituted by a matrix like $\mathbf{W}$ if some elements of $X_t$ are also not observable.

We want to estimate the values of $\mathbf{Y}^o$ given those of $\mathbf{Z}$. We measure the expected loss by the mean squared disaggregation error (MSDE), and formulate the problem as:

$$
\min_{\widetilde{\mathbf{Z}}} tr \left( E(\mathbf{Z}^o - \widetilde{\mathbf{Z}})(\mathbf{Z}^o - \widetilde{\mathbf{Z}})' \right) \quad \text{s.t.} \quad \mathbf{Z} = \overline{\mathbf{W}}\mathbf{Z}^o. \tag{3}
$$

Different weights can be assigned to different errors and cross errors can be taken into account by inserting a symmetric positive semidefinite matrix, $\mathbf{Q}$, into the objective function, thus reformulating the problem as:

$$
\min_{\widetilde{\mathbf{Z}}} tr \left( E(\mathbf{Z}^o - \widetilde{\mathbf{Z}})\mathbf{Q}(\mathbf{Z}^o - \widetilde{\mathbf{Z}})' \right) \quad \text{s.t.} \quad \mathbf{Z} = \overline{\mathbf{W}}\mathbf{Z}^o. \tag{4}
$$

Using the Choleski decomposition $\mathbf{Q} = \mathbf{PP}'$ and defining $\mathbf{R}^o = \mathbf{Z}^o\mathbf{P}^{-1}$, $\mathbf{R} = \mathbf{ZP}^{-1}$, $\widetilde{\mathbf{R}} = \widetilde{\mathbf{Z}}\mathbf{P}^{-1}$, (4) can be written as (3), after substituting $\mathbf{Z}$ with $\mathbf{R}$. Hence, we stick to the formulation in (3) for the objective function to be minimized.

# 3 Estimators and Optimality Results

For the moment we do not assume the factor representation in (1) and (2), but only that second moments of $\mathbf{Z}^o$ exist, and its covariance matrix is

denoted by

$$\underset{(n+1)T\times(n+1)T}{\mathbf{V}_{\mathbf{Z}^\circ}} = \begin{pmatrix} \underset{T\times T}{\mathbf{V}_{\mathbf{Y}^\circ}} & \underset{T\times nT}{\mathbf{C}_{\mathbf{Y}^\circ\mathbf{X}}} \\ \underset{nT\times T}{\mathbf{C}_{\mathbf{X}\mathbf{Y}^\circ}} & \underset{nT\times nT}{\mathbf{V}_{\mathbf{X}}} \end{pmatrix}.$$

This assumption implies the existence of second moments of $\mathbf{Z}$, the observed aggregated variables, whose covariance matrix is

$$\mathbf{V}_{\mathbf{Z}} = \overline{\mathbf{W}}\mathbf{V}_{\mathbf{Z}^\circ}\overline{\mathbf{W}}'.$$

Within this general framework, Proposition 1 characterizes the optimal estimator.

**Proposition 1** *The (linear) minimum MSDE estimator is:*

$$\widehat{\mathbf{Z}} = \mathbf{V}_{\mathbf{Z}^\circ}\overline{\mathbf{W}}'\mathbf{V}_{\mathbf{Z}}^{-1}\mathbf{Z},$$

*with*

$$E(\mathbf{Z}^o - \widehat{\mathbf{Z}})(\mathbf{Z}^o - \widehat{\mathbf{Z}})' = \mathbf{V}_{\mathbf{Z}^\circ} - \mathbf{V}_{\mathbf{Z}^\circ}\overline{\mathbf{W}}'\mathbf{V}_{\mathbf{Z}}^{-1}\overline{\mathbf{W}}\mathbf{V}_{\mathbf{Z}^\circ}.$$

**Proof.** Consider a general linear estimator $\mathbf{PZ} = \mathbf{P}\overline{\mathbf{W}}\mathbf{Z}^o$. The objective function can then be written as

$$tr\left(E(\mathbf{I} - \mathbf{P}\overline{\mathbf{W}})\mathbf{Z}^o\mathbf{Z}^{o\prime}(\mathbf{I} - \mathbf{P}\overline{\mathbf{W}})'\right) = tr\left((\mathbf{I} - \mathbf{P}\overline{\mathbf{W}})\mathbf{V}_{\mathbf{Z}^\circ}(\mathbf{I} - \mathbf{P}\overline{\mathbf{W}})'\right).$$

The optimal projection matrix $\widehat{\mathbf{P}}$ is given by the first order conditions

$$-\mathbf{V}_{\mathbf{Z}^\circ}\overline{\mathbf{W}}' + \widehat{\mathbf{P}}\overline{\mathbf{W}}\mathbf{V}_{\mathbf{Z}}\overline{\mathbf{W}}' = 0.$$

The second order conditions are satisfied for this choice of $\mathbf{P}$, given that $\overline{\mathbf{W}}\mathbf{V}_{\mathbf{Z}^\circ}\overline{\mathbf{W}}'$ is a positive definite matrix. Thus, the linear minimum MSDE estimator is

$$\widehat{\mathbf{Z}} = \widehat{\mathbf{P}}\mathbf{Z} = \mathbf{V}_{\mathbf{Z}^\circ}\overline{\mathbf{W}}'\mathbf{V}_{\mathbf{Z}}^{-1}\mathbf{Z}.$$

Moreover,

$$\begin{aligned}
E(\mathbf{Z}^o - \widehat{\mathbf{Z}})(\mathbf{Z}^o - \widehat{\mathbf{Z}})' &= E(\mathbf{Z}^o - \widehat{\mathbf{P}}\overline{\mathbf{W}}\mathbf{Z}^o)(\mathbf{Z}^o - \widehat{\mathbf{P}}\overline{\mathbf{W}}\mathbf{Z}^o)' \\
&= E\left((\mathbf{I} - \widehat{\mathbf{P}}\overline{\mathbf{W}})\mathbf{Z}^o\mathbf{Z}^{o\prime}(\mathbf{I} - \widehat{\mathbf{P}}\overline{\mathbf{W}})'\right) \\
&= (\mathbf{I} - \widehat{\mathbf{P}}\overline{\mathbf{W}})\mathbf{V}_{\mathbf{Z}^\circ}(\mathbf{I} - \widehat{\mathbf{P}}\overline{\mathbf{W}})' \\
&= ((\mathbf{I} - \widehat{\mathbf{P}}\overline{\mathbf{W}})\mathbf{V}_{\mathbf{Z}^\circ})' - ((\mathbf{I} - \widehat{\mathbf{P}}\overline{\mathbf{W}})\mathbf{V}_{\mathbf{Z}^\circ}\overline{\mathbf{W}}'\widehat{\mathbf{P}}')' \\
&= (\mathbf{V}_{\mathbf{Z}^\circ} - \mathbf{V}_{\mathbf{Z}^\circ}\overline{\mathbf{W}}'\mathbf{V}_{\mathbf{Z}}\overline{\mathbf{W}}\mathbf{V}_{\mathbf{Z}^\circ})' \\
&= \mathbf{V}_{\mathbf{Z}^\circ} - \mathbf{V}_{\mathbf{Z}^\circ}\overline{\mathbf{W}}'\mathbf{V}_{\mathbf{Z}}\overline{\mathbf{W}}\mathbf{V}_{\mathbf{Z}^\circ}. \blacksquare
\end{aligned}$$

Useful insights can be gained by expanding the formula of the optimal predictor as

$$\widehat{\mathbf{Z}} = \left( \begin{array}{c} \widehat{\mathbf{Y}} \\ \widehat{\mathbf{X}} \end{array} \right) = \left( \begin{array}{cc} \underset{T \times s}{\boldsymbol{\alpha}} & \underset{T \times nT}{\boldsymbol{\beta}} \\ \underset{nT \times s}{\boldsymbol{\gamma}} & \underset{nT \times nT}{\boldsymbol{\delta}} \end{array} \right) \left( \begin{array}{c} \mathbf{Y} \\ \mathbf{X} \end{array} \right), \tag{5}$$

where

$$
\begin{aligned}
\boldsymbol{\alpha} &= (\mathbf{V_{Y^o}} - \mathbf{C_{Y^oX}}\mathbf{V_X^{-1}}\mathbf{C_{XY^o}})\mathbf{W}' \left[ \mathbf{W}(\mathbf{V_{Y^o}} - \mathbf{C_{Y^oX}}\mathbf{V_X^{-1}}\mathbf{C_{XY^o}})\mathbf{W}' \right]^{-1}, \\
\boldsymbol{\beta} &= \left[ \mathbf{I} - \mathbf{V_{Y^o}}\mathbf{W}'(\mathbf{W}\mathbf{V_{Y^o}}\mathbf{W}')^{-1}\mathbf{W} \right] \mathbf{C_{Y^oX}} \left[ \mathbf{V_X} - \mathbf{C_{XY^o}}\mathbf{W}'(\mathbf{W}\mathbf{V_{Y^o}}\mathbf{W}')^{-1}\mathbf{W}\mathbf{C_{Y^oX}} \right]^{-1}, \\
\boldsymbol{\gamma} &= \mathbf{0}, \\
\boldsymbol{\delta} &= \mathbf{I}.
\end{aligned}
$$

Clearly, the optimal predictor of $\mathbf{X}$ is $\mathbf{X}$ itself. The matrices $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can instead be interpreted as the coefficients of $\mathbf{Y}$ and $\mathbf{X}$ in a linear projection of $\mathbf{Y}^o$ on $\mathbf{Y}$ and $\mathbf{X}$. In an obvious notation, we have

$$
\begin{aligned}
\boldsymbol{\alpha} &= \mathbf{V_{Y^o|X}}\mathbf{W}'\mathbf{V_{Y|X}^{-1}}, \\
\boldsymbol{\beta} &= \mathbf{V_{Y^o|Y}}\mathbf{W}'\mathbf{V_{X|Y}^{-1}}.
\end{aligned}
$$

We will refer to $\widehat{\mathbf{Z}}$ as the *joint estimator*.

One problem with the joint estimator is that when the dimension of $X_t$ is large, the number of parameters to be estimated is prohibitively large and renders the procedure impossible to implement in practice. This problem can be resolved by imposing sufficient restrictions on the parameters, and the factor representation allows to achieve this goal. Given the factor structure in (1), $X_t$ can be decomposed into a common and an idiosyncratic component, $\Lambda F_t$ and $e_t$, respectively. Stacking $F_t$ and $e_t$ into $\mathbf{F}$ and $\mathbf{e}$, we have,

**Proposition 2** *If* $cov(\mathbf{Y}^o, \mathbf{e} \mid \mathbf{Y}, \mathbf{F}) = 0$, *the optimal estimator is given by:*

$$\widehat{\mathbf{Z}}_F = \left( \begin{array}{cc} \boldsymbol{\alpha}_F & \boldsymbol{\beta}_F \\ \boldsymbol{\gamma}_F & \boldsymbol{\delta}_F \end{array} \right) \left( \begin{array}{c} \mathbf{Y} \\ \mathbf{F} \end{array} \right), \tag{6}$$

*where the dimension of the matrices are as in Proposition 1 but with n=p. In particular,*

$$\boldsymbol{\alpha}_F = (\mathbf{V_{Y^\circ}} - \mathbf{C}_{Y^\circ F}\mathbf{V}_F^{-1}\mathbf{C}_{FY^\circ})\mathbf{W}' \left[\mathbf{W}(\mathbf{V_{Y^\circ}} - \mathbf{C}_{Y^\circ F}\mathbf{V}_F^{-1}\mathbf{C}_{FY^\circ})\mathbf{W}'\right]^{-1},$$

$$\boldsymbol{\beta}_F = \left[\mathbf{I} - \mathbf{V_{Y^\circ}}\mathbf{W}'(\mathbf{W}\mathbf{V_{Y^\circ}}\mathbf{W}')^{-1}\mathbf{W}\right]\mathbf{C}_{Y^\circ F}\left[\mathbf{V}_F - \mathbf{C}_{FY^\circ}\mathbf{W}'(\mathbf{W}\mathbf{V_{Y^\circ}}\mathbf{W}')^{-1}\mathbf{W}\mathbf{C}_{Y^\circ F}\right]^{-1},$$

$$\boldsymbol{\gamma}_F = \mathbf{0},$$

$$\boldsymbol{\delta}_F = \mathbf{I}.$$

*Moreover, $\widehat{\mathbf{Z}}_F$ is more efficient than the joint estimator $\widehat{\mathbf{Z}}$:*

$$E(\mathbf{Z}_F^o - \widehat{\mathbf{Z}}_F)(\mathbf{Z}_F^o - \widehat{\mathbf{Z}}_F)' = \mathbf{V}_{Z_F^o} - \mathbf{V}_{\mathbf{Z}_F^o}\overline{\mathbf{R}}'\mathbf{V}_{\mathbf{Z}_F}^{-1}\overline{\mathbf{R}}\mathbf{V}_{\mathbf{Z}_F^o},$$

*where $\mathbf{Z}_F^o = (\mathbf{Y}^{o'} : \mathbf{F}')'$ , $\mathbf{Z}_F = (\mathbf{Y}' : \mathbf{F}')'$ and $\overline{\mathbf{R}}$ is constructed as $\overline{\mathbf{W}}$ but with n=p.*

**Proof.** When, $\text{cov}(\mathbf{Y}^o, \mathbf{e} \mid \mathbf{Y}, \mathbf{F}) = 0$, the weights in the optimal estimator of $\mathbf{Y}^o$ coincide with those of a projection of $\mathbf{Y}^o$ on $\mathbf{Y}$ and $\mathbf{F}$. In this projection the coefficient of $\mathbf{e}$ is restricted to be zero, which yields the increase in efficiency with respect to $\widehat{\mathbf{Z}}$. ∎

In this case all the relevant information is summarized by the factors. We call $\widehat{\mathbf{Z}}_F$ the *factor estimator.*

If $\beta' = 0$ in (2), so that the factors are uncorrelated with $\mathbf{Y}^o$, an estimator that only exploits the information in the observed data will be more efficient. This is formally stated in the following proposition.

**Proposition 3** *If $\text{cov}(\mathbf{Y}^o, \mathbf{X} \mid \mathbf{Y}) = 0$, the optimal estimator is given by:*

$$\widehat{\mathbf{Z}}_U = \begin{pmatrix} \widehat{\mathbf{Y}}_U \\ \widehat{\mathbf{X}}_U \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha}_U & \boldsymbol{\beta}_U \\ \boldsymbol{\gamma}_U & \boldsymbol{\delta}_U \end{pmatrix} \begin{pmatrix} \mathbf{Y} \\ \mathbf{X} \end{pmatrix}, \tag{7}$$

*where the dimension of the matrices are as in Proposition 1. In particular,*

$$\boldsymbol{\alpha}_U = \mathbf{V_{Y^\circ}}\mathbf{W}'\mathbf{V}_\mathbf{Y}^{-1},$$

$$\boldsymbol{\beta}_U = \mathbf{0},$$

$$\boldsymbol{\gamma}_U = \mathbf{0},$$

$$\boldsymbol{\delta}_U = \mathbf{I}.$$

*Moreover, $\widehat{\mathbf{Z}}_U$ is more efficient than the joint estimator $\widehat{\mathbf{Z}}$, and it is*

$$E(\mathbf{Y}^o - \widehat{\mathbf{Y}}_U)(\mathbf{Y}^o - \widehat{\mathbf{Y}}_U)' = \mathbf{V_{Y^\circ}} - \mathbf{V_{Y^\circ}}\mathbf{W}'\mathbf{V_Y^{-1}}\mathbf{W}\mathbf{V_{Y^\circ}}.$$

**Proof.** When, $cov(\mathbf{Y}^o, \mathbf{X} \mid \mathbf{Y}) = 0$, the weights in the optimal estimator of $\mathbf{Y}^o$ coincide with those of a projection of $\mathbf{Y}^o$ on $\mathbf{Y}$ only. In this projection the coefficient of $\mathbf{X}$ is restricted to be zero, which yields the increase in efficiency with respect to $\widehat{\mathbf{Z}}$. ∎

$\widehat{\mathbf{Z}}_U$ will be called the *univariate estimator*. It is well known and often adopted in the literature, see e.g. Marcellino (1998).

Next, a conditional estimator is defined in

**Proposition 4** *The estimator that solves the problem*

$$\min tr\left(E(\mathbf{Z}^o - \widetilde{\mathbf{Z}})(\mathbf{Z}^o - \widetilde{\mathbf{Z}})' \mid \mathbf{C_{Y^\circ X}V_X X}\right) \quad s.t. \quad \mathbf{Z} = \overline{\mathbf{W}}\mathbf{Z}^o. \tag{8}$$

*is:*

$$\widehat{\mathbf{Z}}_C = \begin{pmatrix} \boldsymbol{\alpha}_C & \boldsymbol{\beta}_C \\ \boldsymbol{\gamma}_C & \boldsymbol{\delta}_C \end{pmatrix} \begin{pmatrix} \mathbf{Y} \\ \mathbf{X} \end{pmatrix}, \tag{9}$$

*where the dimension of the matrices are as in Proposition 1. In particular,*

$$\begin{aligned}
\boldsymbol{\alpha}_C &= (\mathbf{V_{Y^\circ}} - \mathbf{C_{Y^\circ X}V_X^{-1}C_{XY^\circ}})\mathbf{W}'\left[\mathbf{W}(\mathbf{V_{Y^\circ}} - \mathbf{C_{Y^\circ X}V_X^{-1}C_{XY^\circ}})\mathbf{W}'\right]^{-1}, \\
\boldsymbol{\beta}_C &= [\mathbf{I} - \boldsymbol{\alpha}_C]\mathbf{C_{Y^\circ X}V_X}, \\
\boldsymbol{\gamma}_C &= \mathbf{0}, \\
\boldsymbol{\delta}_C &= \mathbf{I}.
\end{aligned}$$

*Moreover, if $cov(\mathbf{Y}, \mathbf{X}) = 0$,*

$$\widehat{\mathbf{Z}}_C = \widehat{\mathbf{Z}}.$$

**Proof.** Define

$$\begin{aligned}
\widetilde{S} &= \mathbf{Y}^o - \mathbf{C_{Y^\circ X}V_X X} \\
\widehat{S} &= \widehat{Y} - \mathbf{C_{Y^\circ X}V_X X} \\
\widetilde{t} &= \mathbf{Y} - \mathbf{W}\mathbf{C_{Y^\circ X}V_X X}.
\end{aligned}$$

The problem can then be reformulated as

$$\min_{\widehat{S}} tr(E(\widetilde{S} - \widehat{S})(\widetilde{S} - \widehat{S})'|\mathbf{C}_{\mathbf{Y}^o\mathbf{X}}\mathbf{V}_{\mathbf{X}}\mathbf{X}) \quad s.t. \quad \widetilde{t} = \mathbf{W}\widetilde{S}. \qquad (10)$$

From Proposition 1, the solution is

$$\widehat{S}^* = V_S W V_t^{-1}\widetilde{t}.$$

Substituting back the expressions for $\widehat{S}$ and $\widetilde{t}$, yields the formula in (9). Under the additional condition $cov(\mathbf{Y},\mathbf{X}) = WC_{YX} = 0$, it is

$$\begin{aligned}
\alpha^C &=& \mathbf{V}_{\mathbf{Y}^o}\mathbf{W}'V_y^{-1} = \alpha; \\
\beta^C &=& \mathbf{C}_{\mathbf{Y}^o\mathbf{X}}\mathbf{V}_{\mathbf{X}} = \beta. \blacksquare
\end{aligned}$$

We call $\widehat{\mathbf{Z}}_C$ the *conditional estimator*. Notice that $\widehat{\mathbf{Y}}_C$ is a convex combination of $\mathbf{Y}$ and $\mathbf{X}$, where the weight on $\mathbf{Y}$ is equal to that in $\widehat{\mathbf{Y}}$ in the joint estimator $\widehat{\mathbf{Z}}$, but the weight on $\mathbf{X}$ is different, unless $\mathbf{Y}$ and $\mathbf{X}$ are uncorrelated. In terms of projections, it is useful to derive $\widehat{\mathbf{Y}}_C$ in two steps. In the first step $\mathbf{Y}^o$ is projected on $\mathbf{X}$. In the second step, the residuals form the first step are projected on their aggregated counterpart. If $\mathbf{Y}$ and $\mathbf{X}$ are uncorrelated, this procedure is equivalent to projecting $\mathbf{Y}^o$ on $\mathbf{Y}$ and $\mathbf{X}$, which generates $\widehat{\mathbf{Y}}$. Otherwise, the results will be different, as shown in (5) and (9).

The formula in (9) can be extended to the case where a generic preliminary estimator is available, $\mathbf{Y}_p^o$, but it does not satisfy the aggregation constraint $\mathbf{Y} = \mathbf{W}\mathbf{Y}_p^o$. In this case the problem is

$$\min_{\widetilde{\mathbf{Z}}} tr\left(E(\mathbf{Z}^o - \widetilde{\mathbf{Z}})(\mathbf{Z}^o - \widetilde{\mathbf{Z}})' \mid \mathbf{Y}_p^o\right) \quad s.t. \quad \mathbf{Z} = \overline{\mathbf{W}}\mathbf{Z}^o, \qquad (11)$$

and it can be easily shown that the optimal estimator of $\mathbf{Y}^o$ is

$$\widehat{\mathbf{Y}}_P = \mathbf{Y}_p^o + \mathbf{V}_{\mathbf{Y}^o}\mathbf{W}'\mathbf{V}_{\mathbf{Y}}^{-1}(\mathbf{Y} - \mathbf{W}\mathbf{Y}_p^o). \qquad (12)$$

We refer to $\widehat{\mathbf{Y}}_P$ as to the *preliminary estimator*. $\widehat{\mathbf{Y}}_P$ boils down to the conditional estimator $\widehat{\mathbf{Y}}_C$ when $\mathbf{Y}_p^o = \mathbf{C}_{\mathbf{Y}^o\mathbf{X}}\mathbf{V}_{\mathbf{X}}\mathbf{X}$. Chow and Lin's (1971) estimator belongs to this class. In their case $\mathbf{Y}_p^o = \widehat{\gamma}_{GLS}\mathbf{X}$, and $\widehat{\gamma}_{GLS}$ is first obtained from a GLS regression of observed aggregated $Y_t$ on $X_t$. As a consequence, this estimator will be in general inefficient with respect to the joint estimator $\widehat{\mathbf{Y}}$ in (5).

More generally, when the restrictions that lead to $\widehat{\mathbf{Y}}_F$, $\widehat{\mathbf{Y}}_U$, and $\widehat{\mathbf{Y}}_C$ are not satisfied, the resulting estimators will be less efficient than $\widehat{\mathbf{Y}}$. We quantify the loss of efficiency in the next proposition, but some additional notation has to be introduced first. Define,

$$
\underset{(n+1)T\times T(p+1)}{\mathbf{A}} = \begin{pmatrix} \mathbf{I}_T & \mathbf{0} & \mathbf{0} & ... & \mathbf{0} \\ \mathbf{0} & \lambda_{11}\mathbf{I}_T & \lambda_{12}\mathbf{I}_T & ... & \lambda_{1p}\mathbf{I}_T \\ ... & & & & \\ \mathbf{0} & \lambda_{n1}\mathbf{I}_T & \lambda_{n2}\mathbf{I}_T & ... & \lambda_{np}\mathbf{I}_T \end{pmatrix}, \quad \underset{(n+1)T\times 1}{\boldsymbol{\varepsilon}} = \begin{pmatrix} \underset{T\times 1}{\mathbf{0}} \\ \underset{nT\times 1}{\mathbf{e}} \end{pmatrix},
$$

where $\lambda_{ij}$ is the $(i,j)$th element in the factor loading matrix $\Lambda$ in equation (1). Thus, $\mathbf{Z}^o = \mathbf{A}\mathbf{Z}_F^o + \boldsymbol{\varepsilon}$. Also, let $\mathbf{a} = (\boldsymbol{\alpha} : \boldsymbol{\beta})$, $\mathbf{a}_F = (\boldsymbol{\alpha}_F : \boldsymbol{\beta}_F)$, $\mathbf{a}_U = (\boldsymbol{\alpha} : \mathbf{0})$, $\mathbf{a}_C = (\boldsymbol{\alpha}_C : \boldsymbol{\beta}_C)$, $\boldsymbol{\Sigma} = E(\mathbf{Y}^o - \widehat{\mathbf{Y}})(\mathbf{Y}^o - \widehat{\mathbf{Y}})'$, $\boldsymbol{\Sigma}_i = E(\mathbf{Y}^o - \widehat{\mathbf{Y}}_i)(\mathbf{Y}^o - \widehat{\mathbf{Y}}_i)'$, $i = F, U, C$. Then,

**Proposition 5** *If* $cov(\mathbf{Y}^o, \mathbf{e} \mid \mathbf{Y}, \mathbf{F}) \neq 0$, $cov(\mathbf{Y}^o, \mathbf{X} \mid \mathbf{Y}) \neq 0$, $cov(\mathbf{Y}, \mathbf{X}) \neq 0$, *we have*

$$
\begin{aligned}
\boldsymbol{\Sigma}_F - \boldsymbol{\Sigma} &= (\mathbf{a}\mathbf{A} - \mathbf{a}_F)\mathbf{V}_{\mathbf{Z}_F}(\mathbf{a}\mathbf{A} - \mathbf{a}_F)' + (\mathbf{a}\mathbf{A} - \mathbf{a}_F)\mathbf{C}_{\mathbf{Z}_F\boldsymbol{\varepsilon}} + \mathbf{C}_{\boldsymbol{\varepsilon}\mathbf{Z}_F}(\mathbf{a}\mathbf{A} - \mathbf{a}_F)' + \mathbf{V}_{\boldsymbol{\varepsilon}}, \\
\boldsymbol{\Sigma}_U - \boldsymbol{\Sigma} &= (\mathbf{a} - \mathbf{a}_U)\mathbf{V}_{\mathbf{Z}^o}(\mathbf{a} - \mathbf{a}_U)', \\
\boldsymbol{\Sigma}_C - \boldsymbol{\Sigma} &= (\mathbf{a} - \mathbf{a}_C)\mathbf{V}_{\mathbf{Z}^o}(\mathbf{a} - \mathbf{a}_C)'.
\end{aligned}
$$

**Proof.** By definition,

$$
\begin{aligned}
\boldsymbol{\Sigma}_F &= E(\mathbf{Y}^o - \widehat{\mathbf{Y}}_F)(\mathbf{Y}^o - \widehat{\mathbf{Y}}_F)' \\
&= E(\mathbf{Y}^o - \widehat{\mathbf{Y}} + \widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_F)(\mathbf{Y}^o - \widehat{\mathbf{Y}} + \widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_F)' \\
&= E(\mathbf{Y}^o - \widehat{\mathbf{Y}})(\mathbf{Y}^o - \widehat{\mathbf{Y}})' + E(\mathbf{a}\mathbf{Z}^o - \mathbf{a}_F\mathbf{Z}_F^o)(\mathbf{a}\mathbf{Z}^o - \mathbf{a}_F\mathbf{Z}_F^o)',
\end{aligned}
$$

where the second equality follows from the lack of correlation between $\mathbf{Y}^o - \widehat{\mathbf{Y}}$ and $\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_F$ because of the optimality of $\widehat{\mathbf{Y}}$. The proof proceeds along the same line for the other estimators. ∎

Table 1 summarizes the estimators.

# 4 Simulation Experiments

In this section we evaluate the relative performance of the alternative disaggregation methods by means of simulation experiments. In particular, with reference to Table 1, we consider two types of factor estimators, two types of univariate estimators, and a conditional/preliminary estimator, while we do not analyze the joint estimator because it is not applicable with a large information set. In the first subsection we provide additional details on these estimators. In the second subsection we describe the design of the experiments. In the final subsection we discuss the results.

## 4.1 Practical Implementation

The practical implementation of the estimators described in the previous section is complicated by two main issues. First of all, in general the variance covariance matrix at the disaggregate level, $\mathbf{V}_{\mathbf{Z}^\circ}$ is not known and has to be derived from its aggregate counterpart, $\mathbf{V}_{\mathbf{Z}}$. This raises a serious identification problem, because several $\mathbf{V}_{\mathbf{Z}^\circ}$ are compatible with $\mathbf{V}_{\mathbf{Z}}$, in the sense that they satisfy the constraint $\mathbf{V}_{\mathbf{Z}} = \overline{\mathbf{W}} \mathbf{V}_{\mathbf{Z}^\circ} \overline{\mathbf{W}}$. Such an issue is often overlooked and it is usual to assume that $\mathbf{V}_{\mathbf{Z}^\circ}$ is known. Marcellino (1998) discusses in more detailes the identification problem when the disaggregated generating mechanism belongs to the ARMA class.

The second issue is estimation of the aggregate variance-covariance matrix, $\mathbf{V}_{\mathbf{Z}}$ or $V_y$. Without making any parametric assumptions on the generating mechanism of the process, estimation of the high order lags of the autocovariance function is highly imprecise in finite samples. Moreover, several elements in these matrices are likely very small or close to zero, which creates an additional problem for the computation of the inverse of the matrices, and for the numerical accuracy of the procedure. Also in this case, assuming a disaggregate ARMA generating mechanism can be helpful.

To take into consideration these two issues, we will experiment with the following estimators.

For the univariate estimator, we assume an AR(3) model at the disaggregate level, and compute the optimal estimator of the missing observations using the Kalman filter, and the smoother, according to the formulae in Harvey and Pierse (1984), see also Kohn and Ansley (1986), Nijman and Palm (1986), and Gomez and Maravall (1994).

As an alternative univariate estimator that does not require assumptions on the dissagregate generating mechanism, we use spline functions, see e.g. Micula and Micula (1998). The tension factor, which indicates the curviness of the resulting function is set equal to one. Values close to zero would imply that the curve is approximately the tensor product of cubic splines, while if the tension factor is large the resulting curve is approximately bi-linear.

To construct a conditional estimator, we use the Chow and Lin (1971) procedure, allowing for an AR(1) structure in the errors of the regression. Five variables are included as regressors, and they are selected among the set of available variables on the basis of their correlation at the aggregate level with the variable to be disaggregated.

Next, we consider two types of factor based estimators. One is based on factors estimated from a balanced panel, i.e., without using information from the variable to be disaggregated. This boils down to applying the Chow and Lin (1971) procedure using (three) estimated factors as regressors rather than some selected variables. The second factor based estimator uses factors extracted from an unbalanced panel, using an EM algorith developed by Stock and Watson (1998). Basically, the disaggregated variable obtained by the first factor method is added to the balanced panel, factors are re-extracted, the Chow and Lin (1971) procedure is applied with the new factors, a new set of disaggregated values are obtained, and they are used to construct another balanced panel, another set of factors, etc. The procedure is repeated until the estimates of the factors do not change substantially in successive iterations. If the fit of the Chow and Lin (1971) regression in the second step is lower than that in the first step, the procedure is stopped and the balanced factor based estimator is used. Following the same line of reasoning as in Stock and Watson (1998) in a forecasting context, the fact that the estimated rather than the true factors are used in the procedure does not affect the quality of the fit of the regression, at least asymptotically, see also Bai (2003).

Finally, it is worth noting that changes in the specification of the estimators under analysis in general do not affect the results substantially.

## 4.2 The experimental design

We consider two different generating mechanisms for the variables:

$$X_t = \Lambda F_t + e_t, \tag{13}$$
$$y_t^o = \beta' F_t + \varepsilon_t,$$

and

$$X_t = Q X_{t-1} + e_t, \tag{14}$$
$$y_t^o = \gamma y_{t-1}^o + \varepsilon_t.$$

The former is a factor model, where the number of factors is set equal to 3, the factors are independent AR(1) processes with root equal to 0.8, and the elements of $\Lambda$ and $\beta$ are independent draws from a uniform distribution over the interval $[0, 1]$. The latter is a set of uncorrelated AR(1) processes, each with root equal to 0.8 ($Q$ is a diagonal matrix). In both cases $e_t$ and $\varepsilon_t$ are i.i.d. $N(0, 1)$ errors, uncorrelated across themselves, $X_t$ contains 50 variables while $y_t^o$ is univariate, and the sample size is set equal to 100.

When the generating mechanism is (13) we expect the factor estimator to be the best, but the Chow and Lin (1971) method should also perform well since the number of regressors (five) is larger than the number of factors, so that the former can provide a good approximation for the latter. When data are generated according to (14) the univariate estimators should be ranked first, since in this case the multivariate methods boil down to simple linear interpolation. The third set of experiments we consider deals with misspecification. We use the factor model to generate the data, but there are ten factors in the DGP while only five are used in the factor based interpolation procedure. Hence, though more complicated models could be used, those in (13) and (14) already provide a good framework to evaluate the relative merits of the alternative interpolation methods.

We set the disaggregation frequency at 4, so that only 25 values of $y_t^o$ can be observed. This mimics disaggregation of annual data into quarterly data. We analyze both stock and flow variables. Next we also consider the case of missing observations at the beginning of the series, assuming that either 5 or 40 starting values of $y_t^o$ are unobservable. For each case we run 2000 replications, and rank the estimators on the basis of the average absolute and mean square disaggregation error (MAE and MSE, respectively).

We also compute percentiles of the distribution of the absolute and mean square disaggregation error, which provides additional information on the performance of the estimators.

## 4.3 Results

The Monte Carlo results, summarized in Tables 2-4, indicate that the MSE and the MAE lead to similar rankings of the various interpolation methods. Moreover, the mean and the median of the distribution of the disaggregation errors are in general very close, with a few exceptions in the case of the Kalman smoother. Hence, in what follows, we focus on the ranking based on the median of the MSE.

A first, robust across experiments, finding is that the balanced panel factor method dominates in a large majority of cases the unbalanced panel approach. This happens for about 70-90% of the replications for most experiments, with lower figures only in the case of the estimation of a low number of missing observations. This is an important finding since it indicates that when more than one series needs to be interpolated (or backdated), it would not be advisable to use the partial information contained into the other series with incomplete coverage to improve the estimates for any given incomplete series, unless very few observations are missing.

When the data are generated by a factor model, the figures in Table 2 clearly show that the factor method performs best. The only exception is the case of an incomplete flow variable, where the other multivariate method, namely the Chow and Lin procedure, yields slightly better results. This may be related to the design of the experiment, since the Chow and Lin regressors are carefully selected on the basis of their correlation properties with the incomplete series.

It is also worth noting that with this DGP the univariate methods do not perform satisfactorily, since neither the Spline, nor the Kalman filter or smoother come close to the multivariate interpolation methods in any of the experiments conducted. The differences are smaller when evaluated on the basis of the MAE, but still the performance is in general 50% to 100% worse.

When the data are generated by independent univariate AR processes, in turn, univariate methods would be expected to provide better estimates, but

the results in Table 3 show that this is not a clear-cut case. For interpolation of stock and flow variables, the Spline method is the best, with the Kalman filter and smoother ranked second, but the factor estimator is a close third best, its MAE performance is only about 10% worse than the parametric univariate methods. In addition, the factor method ranks first in the missing observation case, when 40% of the observations are missing.

The final set of experiments we consider deals with misspecification. In Table 4, a 10-factor model generates the data, but a 5-factor model underlies the interpolation procedure. Notwithstanding this misspecification, the factor method still substantially outperforms the univariate approaches, but the Chow and Lin remains a very valid alternative.

In summary, the factor based method appears to perform quite well in the simulation experiments, even when it is based on a misspecified model. The Chow and Lin approach is ranked a close second, while the univariate methods perform well only with independent processes, which is quite an unlikely situation in practice.

# 5  Applications

In this section we compare the relative merits of the interpolation methods using data for some European countries. In particular, we consider quarterly series for GDP growth and inflation (measured as the quarter on quarter change in the private consumption deflator) for Austria, France, Finland, Germany, Italy, Spain and the Netherlands, over the period 1977:3-1999:2.[1] We carry out two kinds of interpolation exercises. First, we drop all the observations but those corresponding to the last quarter of each year. Second, we drop the initial 20% of the observations. In both cases, we interpolate the missing observations so as to recreate them, and then compare the interpolated with the actual values. The price deflator is treated as a stock variable and GDP growth as a flow.

For inflation, the factors are extracted from a dataset that contains, for all the countries under analysis, several price variables (in growth rates), such as CPI, GDP deflator, export and import deflators, etc., overall 50 series. For GDP growth, we use a set of real variables, that includes among

---

[1] For The Netherlands only GDP growth is analyzed since deflator series are not available over the full sample.

others GDP components, capacity utilization, industrial production, employment and the unemployment rate, etc., a total of 82 series. The two datasets are extracted from the one used in Angelini et al. (2001), and the Data Appendix contains a list of all the series employed in the current analysis. As in the simulation experiments, we extract three factors in each case. Previous work by Stock and Watson (1998) for the US, and Marcellino et al. (2001) and Angelini et al. (2001) for Europe have shown that a limited number of factors are sufficient to explain a substantial proportion of the variability of all the series. We use the same setup as in the simulations also for the Chow and Lin method (namely five regressors are selected from the datasets used for factor extraction, following the procedure outlined in the previous section) and for the univariate methods. The comparison of the methods is based on the mean square and mean absolute disaggregation errors, and all results are summarized in Table 5.

As regards the interpolation of missing infra-year data, in the case of the inflation rates, the Chow and Lin method delivers the best results for 5 of the 6 countries, the only exception being Austria for which the factor procedure works best. In the case of GDP growth, the multivariate methods are again superior, being the best in 5 out of 7 countries. The performance of the factor and Chow and Lin procedures is now similar, with the latter being better than the former in 3 cases (Austria, Germany and Italy), vice versa in 2 cases (Spain and the Netherlands), with a mixed outcome in 2 cases (Finland and France). A similar pattern emerges in the other interpolation exercise, i.e. when estimating missing observations that are concentrated at the beginning of the sample. Multivariate methods are better than univariate methods, Chow and Lin is always the best for the price series, and its performance is similar to the factor based procedure for GDP growth.

To evaluate the robustness of the results we have (a) increased the number of factors to five, as the number of regressors in the Chow and Lin method; (b) decreased the number of regressors in the Chow and Lin method to three, as the number of factors in the base case; (c) used the consumer price index instead of the consumption deflator. Although there were some changes in the resulting figures, the ranking of the interpolation methods was virtually unaltered in all cases.

Overall, these results are in line with the outcome of the simulation experiments and indicate that the gains from using multivariate interpolation

procedures can be substantial, though the traditional Chow and Lin procedure combined with our variable selection strategy is a strong competitor for the new factor based method.

# 6   Using the interpolated data

On the top of the actual-interpolated comparison, which indicates the extent to which the interpolated series fit the actual underlying data, it may be worth assessing the extent to which using the interpolated series instead of the actual ones would impact on possible subsequent econometric exercises. Since the disaggregation error can be considered as a measurement error, we can expect the dynamic properties of the interpolated series and its relationships with other variables to be somewhat affected, with the extent of the bias depending on the goodness of the disaggregation method but also on the specific econometric characteristic under analysis. In particular, in this section we investigate the autocorrelation properties of the interpolated data as well as regression results, both in simulation experiments and using the real data in the previous section.

For the simulations, we generate the data according to the factor model and the AR DGPs in equations (13) and (14). Then we compute the difference ($\rho$) between the first order autocorrelation coefficients for the actual and interpolated series, and the absolute value of the difference ($\beta$) of the estimated coefficient of $x_t$ in the regression $y_t = x_t + u_t$, with $u_t$ i.i.d. N(0,1), using actual and interpolated data for both $y_t$ and $x_t$.

The results are reported in Tables 6 and 7 for the two types of DGPs, and each Table presents figures for stock and flow variables, and for a different fraction of missing observations at the beginning of the sample (either 5% or 40%). As before, we report both the mean and percentiles of the empirical distribution of $\rho$ and $\beta$ over 2000 replications.

Three main comments can be made. First, the ranking of the disaggregation methods in terms of bias reflects that of Tables 2 and 3, which suggests that minimizing the mean square disaggregation error is a good criterion to minimize also the bias in subsequent econometric analyses with the interpolated series. Second, the size of $\rho$ and $\beta$ is much smaller in the case of missing observations at the beginning of the sample than for interpolation of stock and flow variables, which is again in line with the results in Tables

2 and 3 and is mainly due to the lower fraction of missing data, i.e. 5% or 40% versus 75% in the case of stock and flow variables. Third, in general $\beta$ is smaller than $\rho$, indicating that the estimation of dynamic relationships can be more affected by interpolation than contemporaneous relationships, which is also a sensible result.

As far as the application with real data is concerned, we compute $\rho$ as before, while $\beta$ is the difference of the estimated coefficients in a regression of inflation or GDP growth for country $i$ on the same variable for country $j$, using actual and interpolated series. The results are summarized in Table 8 and three main comments are again in order. First, for inflation the lowest values for $\rho$ are achieved by the factor method in 4 out of 6 cases, with Chow and Lin being the best in the remaining two cases. On the other hand, Chow and Lin generates the lowest values for $\beta$ in 3 out of 5 cases, with the spline and the smoother performing best in the other two cases. The biases are in general small, ranging for $\rho$ between 0.001 and 0.12, and for $\beta$ between 0.001 and 0.035. Second, for GDP growth Chow and Lin is the best both in terms of $\rho$ (6 out 7 cases) and of $\beta$ (4 out of 6 cases). The interesting result is that now the biases are larger, in the range 0.02-0.60 for $\rho$ and 0.008-0.23 for $\beta$. This is presumably related to the lowest persistence of GDP growth with respect to the inflation rate. Third, for the case of missing observations at the beginning of the sample Chow and Lin is clearly the best as regards $\beta$ for inflation, while the results are evenly distributed for $\rho$ and for GDP growth. Both biases, for both variables, are substantially smaller than in the case of interpolation.

The even better performance of the Chow and Lin procedure in the empirical analysis with respect to the simulations is likely due to the covariance structure of the datasets, that is such that there exist some variables highly correlated both at the disaggregate and at the aggregate level with the series to be interpolated. In this context, the variable selection procedure implemented for the Chow and Lin method manages to pick up these variables, while the factor method does not take into consideration the correlation with the variable of interest when extracting the factors. On the other hand, the sizeable biases that can emerge in the estimation of the first order autocorrelation function using interpolated data provide a warning for the interpretation of the results of dynamic models estimated with interpolated data.

# 7 Conclusions

In this paper we have developed a factor based approach to interpolation and estimation of missing observations. The method can exploit the information in very large datasets, and hence it is expected to perform better than existing limited information based approaches. We have compared this method with a number of more standard alternative techniques, from a theoretical point of view and using both artificially generated and actual datasets.

First, the theoretical analysis indicates that large information sets are potentially useful, though the resulting estimators are computationally not feasible unless some restrictions are imposed on the generating mechanism of the data, such as a factor structure.

Second, we have run Monte Carlo experiments in which deleted data from artificial series were re-estimated using the whole range of considered methods (Kalman filter and smoother, Spline, Chow and Lin, factor models). Using a sample of 25 years of quarterly data for 50 series, four cases were examined, namely two in which stock and flow variables are only available at the annual frequency, and also two with variables for which there are missing backdata, amounting to 5% or 40% of the whole sample. Experiments were conducted with DGP's being AR(1) or factor models. To allow for some impact of misspecification, we also estimated factor models comprising a number of factors largely inferior to that of the DGP. Performance was evaluated by the Mean Absolute (interpolation / backdating) Error, Mean Squared Error and the quantiles of the absolute or squared difference between the interpolated series and the original "true" one.

The conclusion of the simulation experiments is that with a factor-DGP, factor method tends to dominate all of the others, although the Chow and Lin method also performs well. Univariate methods, on the contrary, yield poor results. When the DGP is univariate, as expected, univariate methods do the best job, in particular the Spline, but the factor method gives comparable results. On the other hand, real-life data is not very likely to follow such a simplistic DGP.

Third, we have used actual time-series, namely quarterly GDP and inflation for 7 countries of the euro area, for which either all observations are dropped but the last quarter each year or 20% of the sample is dropped, at the earlier part of it, thereby mimicking the experimental design employed

for the artificial series.

The results are similar to the factor-DGP Monte Carlo results, with the multivariate methods clearly outperforming the univariate ones. The Chow and Lin technique in particular delivers very good results overall, in particular for inflation. One reason to explain this comparatively better performance, with respect to the factor method, is that the variables to be used in the Chow and Lin procedure were pre-selected according to the correlation with the series to be interpolated / backdated. Although this biases somewhat the experiment against the factor method, such an approach is however supposed to reflect practioners' standard practice.

Finally, we have tried to assess the extent to which using such interpolated series in subsequent econometric exercises could affect the results. This was done also using both artificial and actual series, checking the extent to which substituting the interpolated / backdated series to the original ones would affect both the estimated first order autocorrelation and a regression coefficient between two series. The results this time were more favourable to the Chow and Lin technique, in particular for growth. This presumably stresses again the importance of the pre-selection of most appropriate variables before running the interpolation procedure. An interesting caveat resulting from the analysis is that biases can be sizeable, especially in the case of interpolation where there are a relatively large number of missing observations.

# References

[1] Angelini, E., Henry, J. and R. Mestre (2001), "Diffusion index based inflation forecasts for the Euro area", European Central Bank WP 61.

[2] Artis, M., Banerjee, A. and M. Marcellino (2001), "Factor forecasts for the UK", CEPR WP 3119.

[3] Bai, J. (2003), "Inferential theory for factor models of large dimension", *Econometrica*, 71, 135-171.

[4] Boot, J.C.G., Feibes, W. and J. H. C. Lisman (1967), "Further Methods of Derivation of Quarterly Figures from Annual Data." *Applied Statistics*, 16, 65-75.

[5] Chamberlain, G. and M. Rothschild (1983), "Arbitrage factor structure, and mean variance analysis of large asset markets", *Econometrica*, 51, 1281-1304.

[6] Chan, W.S. (1993), "Disaggregation of Annual Time-Series Data into Quarterly Figures: A Comparative Study." *Journal of Forecasting*, 12, 677-688.

[7] Chow, G.C. and Lin, A. (1971), "Best Linear Unbiased Interpolation, Distribution and Extrapolation of Time-Series by Related Series." *Review of Economics and Statistics*, 53, 372-375.

[8] Cohen, K.J., Müller, W. and M. W. Padberg (1971), "Autoregressive Approaches to Disaggregation of Time Series Data." *Applied Statistics*, 20, 119-129.

[9] Connor, G. and R.A. Korajczyk (1986), "Performance measurement with the arbitrage pricing theory", *Journal of Financial Economics*, 15, 373-394.

[10] Connor, G. and R.A. Korajczyk (1993), "A test for the number of factors in an approximate factor model", *Journal of Finance*, 48, 1263-1291.

[11] Denton, F. (1971). "Adjustment of Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization." *Journal of the American Statistical Association*, 66, 99-101.

[12] Fernandez, R.B. (1981), "A Methodological Note on the Estimation of Time-Series." *Review of Economics and Statistics*, 63, 471-476.

[13] Forni, M., Hallin, M., Lippi, M. and L. Reichlin (2000), "The generalized factor model: identification and estimation", *The Review of Economic and Statistics*, 82, 540-554.

[14] Gomez, V. and A. Maravall (1994), "Estimation, Prediction and Interpolation for Non Stationary Time Series with the Kalman Filter." *Journal of the American Statistical Association*, 89, 611-624.

[15] Guerrero, V.M. (1990), "Temporal Disaggregation of Time-Series: An ARIMA-Based Approach." *International Statistical Review*, 58, 29-46.

[16] Harvey, A.C. and R.G. Pierse (1984), "Estimating Missing Observations in Economic Time Series." *Journal of the American Statistical Association*, 79, 125-131.

[17] Kohn, R. and C.F. Ansley (1986), "Estimation, Prediction and Interpolation for ARIMA models with Missing Data." *Journal of the American Statistical Association*, 81, 751-761.

[18] Lisman J.H.C. and J. Sandee (1964), "Derivation of Quarterly Figures from Annual Data." *Applied Statistics*, 13, 87-90.

[19] Litterman, R.B. (1983), "A Random Walk, Markov Model for the Distribution of Time Series." *Journal of Business and Economic Statistics*, 1, 169-173.

[20] Marcellino, M. (1998), "Temporal disaggregation, missing observations, outliers, and forecasting: a unifying non-model based procedure", *Advances in Econometrics*, 13, 181-202.

[21] Marcellino, M., Stock, J.H. and M. W. Watson (2001), "Macroeconomic forecasting in the Euro area: country specific versus euro wide information", *European Economic Review*, (forthcoming).

[22] Micula, G. and S. Micula (1998), *Handbook of Splines*, Dordrecht: Kluwer Academic Publishers.

[23] Nijman, T.E. and F. C. Palm (1986), The Construction and Use of Approximations for Missing Quarterly Observations: A Model-Based Approach." *Journal of Business and Economic Statistics*, 4, 47-58.

[24] Stock, J.H. and M.W. Watson (1998), "Diffusion indexes", NBER WP 6702.

[25] Stram, D.O. and W.W.S. Wei (1986), "A Methodological Note on Disaggregation of Time Series Totals." *Journal of Time Series Analysis*, 7, 293-302.

[26] Wei, W.W.S. and D. O. Stram (1990). "Disaggregation of Time Series Models." *Journal of the Royal Statistical Society*, Series B, 52, 453-467.

**Table 1. Alternative Estimators**

$$Joint: \quad \widehat{\mathbf{Y}} = \boldsymbol{\alpha}\mathbf{Y} + \boldsymbol{\beta}\mathbf{X}$$

$$\boldsymbol{\alpha} = (\mathbf{V_{Y^o}} - \mathbf{C_{Y^oX}}\mathbf{V_X^{-1}}\mathbf{C_{XY^o}})\mathbf{W}' \left[\mathbf{W}(\mathbf{V_{Y^o}} - \mathbf{C_{Y^oX}}\mathbf{V_X^{-1}}\mathbf{C_{XY^o}})\mathbf{W}'\right]^{-1}$$

$$\boldsymbol{\beta} = \left[\mathbf{I} - \mathbf{V_{Y^o}}\mathbf{W}'(\mathbf{W}\mathbf{V_{Y^o}}\mathbf{W}')^{-1}\mathbf{W}\right]\mathbf{C_{Y^oX}}\left[\mathbf{V_X} - \mathbf{C_{XY^o}}\mathbf{W}'(\mathbf{W}\mathbf{V_{Y^o}}\mathbf{W}')^{-1}\mathbf{W}\mathbf{C_{Y^oX}}\right]^{-1}$$

$$Factor: \quad \widehat{\mathbf{Y}}_F = \boldsymbol{\alpha}_F\mathbf{Y} + \boldsymbol{\beta}_F\mathbf{F}$$

$$\boldsymbol{\alpha}_F = (\mathbf{V_{Y^o}} - \mathbf{C}_{Y^oF}\mathbf{V}_F^{-1}\mathbf{C}_{FY^o})\mathbf{W}' \left[\mathbf{W}(\mathbf{V_{Y^o}} - \mathbf{C}_{Y^oF}\mathbf{V}_F^{-1}\mathbf{C}_{FY^o})\mathbf{W}'\right]^{-1}$$

$$\boldsymbol{\beta}_F = \left[\mathbf{I} - \mathbf{V_{Y^o}}\mathbf{W}'(\mathbf{W}\mathbf{V_{Y^o}}\mathbf{W}')^{-1}\mathbf{W}\right]\mathbf{C}_{Y^oF}\left[\mathbf{V}_F - \mathbf{C}_{FY^o}\mathbf{W}'(\mathbf{W}\mathbf{V_{Y^o}}\mathbf{W}')^{-1}\mathbf{W}\mathbf{C}_{Y^oF}\right]^{-1}$$

$$Univariate: \quad \widehat{\mathbf{Y}}_U = \boldsymbol{\alpha}_U\mathbf{Y}$$

$$\boldsymbol{\alpha}_U = \mathbf{V_{Y^o}}\mathbf{W}'\mathbf{V_Y^{-1}}$$

$$Conditional: \quad \widehat{\mathbf{Y}}_C = \boldsymbol{\alpha}_C\mathbf{Y} + \boldsymbol{\beta}_C\mathbf{X}$$

$$\boldsymbol{\alpha}_C = (\mathbf{V_{Y^o}} - \mathbf{C_{Y^oX}}\mathbf{V_X^{-1}}\mathbf{C_{XY^o}})\mathbf{W}' \left[\mathbf{W}(\mathbf{V_{Y^o}} - \mathbf{C_{Y^oX}}\mathbf{V_X^{-1}}\mathbf{C_{XY^o}})\mathbf{W}'\right]^{-1}$$

$$\boldsymbol{\beta}_C = [\mathbf{I} - \boldsymbol{\alpha}_C]\mathbf{C_{Y^oX}}\mathbf{V_X}$$

$$Preliminary: \quad \widehat{\mathbf{Y}}_P = \mathbf{Y}_p^o + \boldsymbol{\alpha}_U(\mathbf{Y} - \mathbf{W}\mathbf{Y}_p^o)$$

$$\boldsymbol{\alpha}_U = \mathbf{V_{Y^o}}\mathbf{W}'\mathbf{V_Y^{-1}}$$

Note: See Section 2 for a definition of the relevant matrices.

## Table 2. Disaggregation error, DGP DFM 3 factors

### STOCK

| | MSE | | | | | | MAE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | avg | .05 | .25 | .50 | .75 | .95 | avg | .05 | .25 | .50 | .75 | .95 |
| **DFM** | 0.306 | 0.130 | 0.200 | 0.266 | 0.372 | 0.613 | 0.373 | 0.251 | 0.310 | 0.357 | 0.423 | 0.543 |
| **Chow − Lin** | 0.381 | 0.175 | 0.264 | 0.342 | 0.459 | 0.706 | 0.418 | 0.291 | 0.356 | 0.405 | 0.473 | 0.582 |
| **Spline** | 1.173 | 0.720 | 0.981 | 1.170 | 1.366 | 1.624 | 0.738 | 0.583 | 0.676 | 0.741 | 0.803 | 0.878 |
| **K − filter** | 0.859 | 0.605 | 0.737 | 0.837 | 0.947 | 1.176 | 0.639 | 0.537 | 0.593 | 0.634 | 0.676 | 0.757 |
| **K − smoother** | 1.403 | 0.593 | 0.739 | 0.840 | 0.955 | 1.224 | 0.648 | 0.532 | 0.593 | 0.634 | 0.677 | 0.764 |

Fraction of cases where balanced panel works better than non balanced panel: 0.903

### FLOW

| | MSE | | | | | | MAE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | avg | .05 | .25 | .50 | .75 | .95 | avg | .05 | .25 | .50 | .75 | .95 |
| **DFM** | 0.463 | 0.249 | 0.364 | 0.450 | 0.556 | 0.707 | 0.537 | 0.396 | 0.479 | 0.537 | 0.595 | 0.673 |
| **Chow − Lin** | 0.359 | 0.164 | 0.246 | 0.326 | 0.443 | 0.667 | 0.470 | 0.323 | 0.399 | 0.458 | 0.532 | 0.656 |
| **Spline** | 0.621 | 0.402 | 0.528 | 0.629 | 0.719 | 0.818 | 0.628 | 0.504 | 0.580 | 0.636 | 0.679 | 0.730 |
| **K − filter** | 0.653 | 0.433 | 0.560 | 0.652 | 0.733 | 0.829 | 0.641 | 0.524 | 0.597 | 0.647 | 0.685 | 0.735 |
| **K − smoother** | 0.645 | 0.427 | 0.557 | 0.650 | 0.733 | 0.828 | 0.639 | 0.520 | 0.594 | 0.646 | 0.685 | 0.736 |

Fraction of cases where balanced panel works better than non balanced panel: 0.710

### MISSING OBSERVATIONS 40%

| | MSE | | | | | | MAE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | avg | .05 | .25 | .50 | .75 | .95 | avg | .05 | .25 | .50 | .75 | .95 |
| **DFM** | 0.150 | 0.063 | 0.096 | 0.133 | 0.184 | 0.296 | 0.191 | 0.127 | 0.157 | 0.184 | 0.219 | 0.275 |
| **Chow − Lin** | 0.173 | 0.077 | 0.115 | 0.157 | 0.214 | 0.321 | 0.206 | 0.139 | 0.172 | 0.200 | 0.236 | 0.291 |
| **K − smoother** | 0.814 | 0.217 | 0.364 | 0.441 | 0.534 | 0.940 | 0.375 | 0.264 | 0.306 | 0.339 | 0.375 | 0.518 |

Fraction of cases where balanced panel works better than non balanced panel: 0.730

### MISSING OBSERVATIONS 5%

| | MSE | | | | | | MAE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | avg | .05 | .25 | .50 | .75 | .95 | avg | .05 | .25 | .50 | .75 | .95 |
| **DFM** | 0.018 | 0.003 | 0.008 | 0.014 | 0.024 | 0.047 | 0.023 | 0.010 | 0.016 | 0.022 | 0.029 | 0.041 |
| **Chow − Lin** | 0.021 | 0.004 | 0.010 | 0.017 | 0.028 | 0.053 | 0.025 | 0.012 | 0.018 | 0.024 | 0.031 | 0.044 |
| **K − smoother** | 0.047 | 0.009 | 0.023 | 0.040 | 0.063 | 0.111 | 0.039 | 0.017 | 0.029 | 0.037 | 0.047 | 0.065 |

Fraction of cases where balanced panel works better than non balanced panel: 0.586

Note: The table reports the mean and percentiles of the empirical distribution of the MSE and MAE, computed over 2000 replications, when the DGP is as in (13), for different disaggregation methods, types of variables and of missing observations.

Table 3. Disaggregation error, DGP AR(1)

| | **MSE** | | | | | | **MAE** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **avg** | **.05** | **.25** | **.50** | **.75** | **.95** | **avg** | **.05** | **.25** | **.50** | **.75** | **.95** |

**STOCK**

| | **MSE** | | | | | | **MAE** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **avg** | **.05** | **.25** | **.50** | **.75** | **.95** | **avg** | **.05** | **.25** | **.50** | **.75** | **.95** |
| **DFM** | 0.773 | 0.585 | 0.702 | 0.763 | 0.833 | 0.981 | 0.653 | 0.536 | 0.588 | 0.625 | 0.691 | 0.855 |
| **Chow − Lin** | 0.843 | 0.531 | 0.695 | 0.818 | 0.961 | 1.239 | 0.691 | 0.504 | 0.589 | 0.650 | 0.739 | 1.007 |
| **Spline** | 0.545 | 0.292 | 0.406 | 0.511 | 0.645 | 0.889 | 0.502 | 0.343 | 0.427 | 0.487 | 0.557 | 0.709 |
| **K − filter** | 0.717 | 0.376 | 0.533 | 0.679 | 0.831 | 1.164 | 0.613 | 0.404 | 0.505 | 0.581 | 0.665 | 0.940 |
| **K − smoother** | 0.901 | 0.306 | 0.465 | 0.623 | 0.800 | 1.172 | 0.598 | 0.364 | 0.469 | 0.553 | 0.644 | 0.929 |

Fraction of cases where balanced panel works better than non balanced panel: 0.951

**FLOW**

| | **MSE** | | | | | | **MAE** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **avg** | **.05** | **.25** | **.50** | **.75** | **.95** | **avg** | **.05** | **.25** | **.50** | **.75** | **.95** |
| **DFM** | 0.442 | 0.313 | 0.379 | 0.437 | 0.495 | 0.600 | 0.518 | 0.434 | 0.480 | 0.516 | 0.553 | 0.612 |
| **Chow − Lin** | 1.013 | 0.411 | 0.647 | 0.892 | 1.237 | 1.978 | 0.781 | 0.513 | 0.639 | 0.755 | 0.891 | 1.134 |
| **Spline** | 0.240 | 0.132 | 0.183 | 0.228 | 0.286 | 0.390 | 0.380 | 0.287 | 0.337 | 0.377 | 0.420 | 0.487 |
| **K − filter** | 0.382 | 0.186 | 0.273 | 0.354 | 0.462 | 0.630 | 0.474 | 0.341 | 0.411 | 0.464 | 0.529 | 0.623 |
| **K − smoother** | 1.294 | 0.170 | 0.252 | 0.333 | 0.446 | 0.625 | 0.470 | 0.325 | 0.396 | 0.452 | 0.520 | 0.620 |

Fraction of cases where balanced panel works better than non balanced panel: 0.946

**MISSING OBSERVATIONS 40%**

| | **MSE** | | | | | | **MAE** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **avg** | **.05** | **.25** | **.50** | **.75** | **.95** | **avg** | **.05** | **.25** | **.50** | **.75** | **.95** |
| **DFM** | 0.394 | 0.211 | 0.308 | 0.386 | 0.469 | 0.601 | 0.319 | 0.234 | 0.282 | 0.319 | 0.354 | 0.406 |
| **Chow − Lin** | 0.446 | 0.263 | 0.360 | 0.441 | 0.521 | 0.648 | 0.340 | 0.261 | 0.305 | 0.341 | 0.371 | 0.420 |
| **K − smoother** | 1.578 | 0.228 | 0.353 | 0.463 | 0.601 | 0.927 | 0.382 | 0.242 | 0.304 | 0.352 | 0.405 | 0.517 |

Fraction of cases where balanced panel works better than non balanced panel: 0.871

**MISSING OBSERVATIONS 5%**

| | **MSE** | | | | | | **MAE** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **avg** | **.05** | **.25** | **.50** | **.75** | **.95** | **avg** | **.05** | **.25** | **.50** | **.75** | **.95** |
| **DFM** | 0.057 | 0.007 | 0.020 | 0.041 | 0.077 | 0.171 | 0.043 | 0.015 | 0.027 | 0.039 | 0.055 | 0.087 |
| **Chow − Lin** | 0.059 | 0.008 | 0.022 | 0.042 | 0.078 | 0.161 | 0.044 | 0.017 | 0.028 | 0.040 | 0.056 | 0.084 |
| **K − smoother** | 0.042 | 0.005 | 0.014 | 0.029 | 0.055 | 0.128 | 0.036 | 0.013 | 0.022 | 0.033 | 0.045 | 0.074 |

Fraction of cases where balanced panel works better than non balanced panel: 0.815

Note: The table reports the mean and percentiles of the empirical distribution of the MSE and MAE, computed over 2000 replications, when the DGP is as in (14), for different disaggregation methods, types of variables and of missing observations.

**Table 4. Disaggregation error, DGP DFM Mis-specified**

STOCK

| | MSE | | | | | | MAE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **avg** | **.05** | **.25** | **.50** | **.75** | **.95** | **avg** | **.05** | **.25** | **.50** | **.75** | **.95** |
| **DFM** | 0.213 | 0.098 | 0.148 | 0.192 | 0.259 | 0.397 | 0.313 | 0.216 | 0.266 | 0.304 | 0.355 | 0.437 |
| **Chow − Lin** | 0.223 | 0.110 | 0.162 | 0.209 | 0.268 | 0.384 | 0.322 | 0.229 | 0.278 | 0.316 | 0.359 | 0.433 |
| **Spline** | 1.417 | 1.023 | 1.232 | 1.399 | 1.579 | 1.883 | 0.817 | 0.690 | 0.768 | 0.816 | 0.868 | 0.942 |
| **K − filter** | 0.962 | 0.715 | 0.824 | 0.919 | 1.049 | 1.347 | 0.677 | 0.579 | 0.627 | 0.665 | 0.713 | 0.816 |
| **K − smoother** | 1.627 | 0.722 | 0.833 | 0.928 | 1.057 | 1.392 | 0.690 | 0.581 | 0.631 | 0.669 | 0.718 | 0.822 |

Fraction of cases where balanced panel works better than non balanced panel: 0.996

FLOW

| | MSE | | | | | | MAE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **avg** | **.05** | **.25** | **.50** | **.75** | **.95** | **avg** | **.05** | **.25** | **.50** | **.75** | **.95** |
| **DFM** | 0.585 | 0.362 | 0.480 | 0.577 | 0.682 | 0.834 | 0.607 | 0.475 | 0.553 | 0.608 | 0.663 | 0.732 |
| **Chow − Lin** | 0.236 | 0.117 | 0.170 | 0.220 | 0.286 | 0.409 | 0.382 | 0.274 | 0.329 | 0.375 | 0.428 | 0.514 |
| **Spline** | 0.762 | 0.581 | 0.694 | 0.772 | 0.836 | 0.910 | 0.697 | 0.609 | 0.664 | 0.702 | 0.733 | 0.773 |
| **K − filter** | 0.758 | 0.571 | 0.692 | 0.758 | 0.822 | 0.899 | 0.695 | 0.604 | 0.662 | 0.695 | 0.728 | 0.769 |
| **K − smoother** | 4.042 | 0.569 | 0.692 | 0.758 | 0.823 | 0.900 | 0.714 | 0.605 | 0.662 | 0.696 | 0.729 | 0.769 |

Fraction of cases where balanced panel works better than non balanced panel: 0.975

MISSING OBSERVATIONS 40%

| | MSE | | | | | | MAE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **avg** | **.05** | **.25** | **.50** | **.75** | **.95** | **avg** | **.05** | **.25** | **.50** | **.75** | **.95** |
| **DFM** | 0.098 | 0.043 | 0.067 | 0.091 | 0.120 | 0.176 | 0.155 | 0.103 | 0.131 | 0.152 | 0.176 | 0.214 |
| **Chow − Lin** | 0.096 | 0.045 | 0.068 | 0.090 | 0.117 | 0.167 | 0.155 | 0.107 | 0.132 | 0.152 | 0.174 | 0.211 |
| **K − smoother** | 1.170 | 0.296 | 0.377 | 0.443 | 0.529 | 1.422 | 0.396 | 0.273 | 0.310 | 0.340 | 0.374 | 0.600 |

Fraction of cases where balanced panel works better than non balanced panel: 0.997

MISSING OBSERVATIONS 5%

| | MSE | | | | | | MAE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **avg** | **.05** | **.25** | **.50** | **.75** | **.95** | **avg** | **.05** | **.25** | **.50** | **.75** | **.95** |
| **DFM** | 0.012 | 0.002 | 0.006 | 0.010 | 0.016 | 0.028 | 0.019 | 0.009 | 0.014 | 0.018 | 0.024 | 0.033 |
| **Chow − Lin** | 0.012 | 0.002 | 0.006 | 0.009 | 0.015 | 0.028 | 0.019 | 0.009 | 0.014 | 0.018 | 0.023 | 0.032 |
| **K − smoother** | 0.052 | 0.012 | 0.029 | 0.047 | 0.069 | 0.114 | 0.041 | 0.020 | 0.031 | 0.040 | 0.050 | 0.065 |

Fraction of cases where balanced panel works better than non balanced panel: 0.974

Note: The table reports the mean and percentiles of the empirical distribution of the MSE and MAE, computed over 2000 replications, when the DGP is as in (13) but with 10 factors in the DGP and 5 used in the factor model.

<div align="center">**Table 5. Estimation of quarterly data**</div>

<div align="center">INFLATION</div>

| | MSE | | | | | | MAE | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | AT | DE | ES | FI | FR | IT | AT | DE | ES | FI | FR | IT |
| **DFM** | 0.42 | 0.47 | 0.13 | 0.30 | 0.09 | 0.06 | 0.43 | 0.46 | 0.26 | 0.36 | 0.19 | 0.16 |
| **Chow − Lin** | 0.47 | 0.25 | 0.11 | 0.23 | 0.06 | 0.02 | 0.44 | 0.32 | 0.20 | 0.33 | 0.15 | 0.10 |
| **Spline** | 0.55 | 0.70 | 0.28 | 0.34 | 0.18 | 0.07 | 0.49 | 0.55 | 0.33 | 0.39 | 0.27 | 0.18 |
| **K − filter** | 0.57 | 0.60 | 0.34 | 0.50 | 0.18 | 0.12 | 0.50 | 0.49 | 0.40 | 0.46 | 0.31 | 0.24 |
| **K − smoother** | 0.54 | 0.58 | 0.30 | 0.50 | 0.17 | 0.07 | 0.47 | 0.48 | 0.37 | 0.46 | 0.30 | 0.17 |

<div align="center">MISSING OBSERVATIONS 20%</div>

| | MSE | | | | | | MAE | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | AT | DE | ES | FI | FR | IT | AT | DE | ES | FI | FR | IT |
| **DFM** | 0.12 | 0.12 | 0.08 | 0.07 | 0.06 | 0.03 | 0.12 | 0.12 | 0.10 | 0.09 | 0.10 | 0.06 |
| **Chow − Lin** | 0.07 | 0.11 | 0.04 | 0.05 | 0.04 | 0.009 | 0.09 | 0.12 | 0.07 | 0.08 | 0.07 | 0.04 |
| **K − smoother** | 0.15 | 0.46 | 0.17 | 0.14 | 0.71 | 0.30 | 0.15 | 0.25 | 0.16 | 0.15 | 0.30 | 0.18 |

<div align="center">REAL GDP GROWTH</div>

| | MSE | | | | | | | MAE | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | AT | DE | ES | FI | FR | IT | NL | AT | DE | ES | FI | FR | IT | NL |
| **DFM** | 0.80 | 0.75 | 0.36 | 0.81 | 0.40 | 0.56 | 0.57 | 0.64 | 0.66 | 0.45 | 0.70 | 0.53 | 0.57 | 0.55 |
| **Chow − Lin** | 0.74 | 0.53 | 0.43 | 0.81 | 0.40 | 0.39 | 0.73 | 0.63 | 0.57 | 0.53 | 0.70 | 0.50 | 0.48 | 0.63 |
| **Spline** | 0.87 | 0.84 | 0.27 | 0.76 | 0.46 | 0.57 | 0.71 | 0.67 | 0.68 | 0.40 | 0.67 | 0.56 | 0.57 | 0.60 |
| **K − filter** | 0.76 | 0.80 | 0.26 | 0.83 | 0.48 | 0.63 | 0.58 | 0.62 | 0.69 | 0.40 | 0.71 | 0.59 | 0.62 | 0.55 |
| **K − smoother** | 0.78 | 0.86 | 0.28 | 0.79 | 0.50 | 0.63 | 0.58 | 0.63 | 0.71 | 0.41 | 0.69 | 0.61 | 0.62 | 0.55 |

<div align="center">MISSING OBSERVATIONS 20%</div>

| | MSE | | | | | | | MAE | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | AT | DE | ES | FI | FR | IT | NL | AT | DE | ES | FI | FR | IT | NL |
| **DFM** | 0.36 | 0.12 | 0.29 | 0.34 | 0.17 | 0.27 | 0.27 | 0.18 | 0.12 | 0.20 | 0.19 | 0.16 | 0.19 | 0.17 |
| **Chow − Lin** | 0.37 | 0.11 | 0.20 | 0.22 | 0.20 | 0.22 | 0.69 | 0.19 | 0.12 | 0.17 | 0.17 | 0.16 | 0.16 | 0.30 |
| **K − smoother** | 0.40 | 0.30 | 0.20 | 0.26 | 0.30 | 0.42 | 0.35 | 0.20 | 0.20 | 0.16 | 0.19 | 0.19 | 0.24 | 0.20 |

Note: Inflation is treated as a stock variable, GDP growth as a flow variable.

AT: Austria, DE: Germany, ES: Spain, FI: Finland, FR: France, IT: Italy, NL: The Netherlands

## Table 6. Properties of interpolated data, DGP DFM 3 factors

### STOCK

| | ρ | | | | | | β | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | avg | .05 | .25 | .50 | .75 | .95 | avg | .05 | .25 | .50 | .75 | .95 |
| **DFM** | 0.099 | 0.007 | 0.037 | 0.080 | 0.142 | 0.245 | 0.124 | 0.010 | 0.049 | 0.102 | 0.173 | 0.326 |
| **Chow − Lin** | 0.110 | 0.018 | 0.041 | 0.090 | 0.152 | 0.290 | 0.121 | 0.009 | 0.049 | 0.100 | 0.169 | 0.304 |
| **Spline** | 0.670 | 0.303 | 0.514 | 0.671 | 0.820 | 1.038 | 0.111 | 0.009 | 0.044 | 0.092 | 0.157 | 0.277 |
| **K − filter** | 0.335 | 0.027 | 0.138 | 0.290 | 0.494 | 0.802 | 0.168 | 0.014 | 0.068 | 0.145 | 0.242 | 0.404 |
| **K − smoother** | 0.375 | 0.033 | 0.168 | 0.335 | 0.544 | 0.870 | 0.183 | 0.015 | 0.072 | 0.154 | 0.254 | 0.428 |

### FLOW

| | ρ | | | | | | β | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | avg | .05 | .25 | .50 | .75 | .95 | avg | .05 | .25 | .50 | .75 | .95 |
| **DFM** | 0.476 | 0.220 | 0.360 | 0.472 | 0.577 | 0.746 | 0.076 | 0.007 | 0.029 | 0.063 | 0.110 | 0.195 |
| **Chow − Lin** | 0.144 | 0.012 | 0.056 | 0.119 | 0.201 | 0.368 | 0.094 | 0.006 | 0.034 | 0.077 | 0.134 | 0.244 |
| **Spline** | 0.781 | 0.466 | 0.636 | 0.783 | 0.921 | 1.102 | 0.076 | 0.006 | 0.030 | 0.062 | 0.108 | 0.191 |
| **K − filter** | 0.586 | 0.231 | 0.440 | 0.588 | 0.735 | 0.924 | 0.084 | 0.006 | 0.032 | 0.068 | 0.118 | 0.212 |
| **K − smoother** | 0.606 | 0.253 | 0.459 | 0.609 | 0.758 | 0.952 | 0.097 | 0.006 | 0.033 | 0.067 | 0.119 | 0.215 |

### MISSING OBSERVATIONS 40%

| | ρ | | | | | | β | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | avg | .05 | .25 | .50 | .75 | .95 | avg | .05 | .25 | .50 | .75 | .95 |
| **DFM** | 0.051 | 0.004 | 0.018 | 0.041 | 0.072 | 0.136 | 0.053 | 0.004 | 0.020 | 0.043 | 0.074 | 0.137 |
| **Chow − Lin** | 0.054 | 0.004 | 0.019 | 0.043 | 0.076 | 0.140 | 0.052 | 0.004 | 0.020 | 0.041 | 0.072 | 0.140 |
| **K − smoother** | 0.125 | 0.007 | 0.037 | 0.083 | 0.153 | 0.415 | 0.099 | 0.004 | 0.023 | 0.048 | 0.090 | 0.417 |

### MISSING OBSERVATIONS 5%

| | ρ | | | | | | β | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | avg | .05 | .25 | .50 | .75 | .95 | avg | .05 | .25 | .50 | .75 | .95 |
| **DFM** | 0.013 | 0.001 | 0.004 | 0.010 | 0.019 | 0.037 | 0.013 | 0.001 | 0.004 | 0.010 | 0.017 | 0.035 |
| **Chow − Lin** | 0.014 | 0.001 | 0.005 | 0.011 | 0.019 | 0.039 | 0.013 | 0.001 | 0.005 | 0.010 | 0.018 | 0.034 |
| **K − smoother** | 0.021 | 0.001 | 0.006 | 0.014 | 0.028 | 0.064 | 0.014 | 0.001 | 0.005 | 0.011 | 0.019 | 0.040 |

Note: The table reports the difference ($\rho$) between the first order autocorrelation coefficients for the actual and interpolated series, and the absolute value of the difference ($\beta$) of the estimated coefficient of $x_t$ in the regression $y_t = x_t + u_t$, with $u_t$ i.i.d. N(0,1), using actual and interpolated data for both $y_t$ and $x_t$. The DGP is as in (13).

## Table 7. Properties of interpolated data, DGP AR(1)

| | | STOCK | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\rho$ | | | | | | $\beta$ | | | | | |
| | avg | .05 | .25 | .50 | .75 | .95 | avg | .05 | .25 | .50 | .75 | .95 |
| **DFM** | 0.615 | 0.243 | 0.503 | 0.648 | 0.752 | 0.876 | 0.137 | 0.009 | 0.051 | 0.115 | 0.194 | 0.346 |
| **Chow − Lin** | 0.463 | 0.154 | 0.319 | 0.450 | 0.602 | 0.809 | 0.187 | 0.016 | 0.079 | 0.166 | 0.273 | 0.437 |
| **Spline** | 0.091 | 0.009 | 0.045 | 0.084 | 0.127 | 0.202 | 0.107 | 0.009 | 0.042 | 0.088 | 0.151 | 0.266 |
| **K − filter** | 0.302 | 0.019 | 0.110 | 0.276 | 0.448 | 0.665 | 0.181 | 0.012 | 0.069 | 0.151 | 0.268 | 0.447 |
| **K − smoother** | 0.247 | 0.016 | 0.087 | 0.187 | 0.366 | 0.614 | 0.199 | 0.014 | 0.077 | 0.163 | 0.288 | 0.473 |

| | | FLOW | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\rho$ | | | | | | $\beta$ | | | | | |
| | avg | .05 | .25 | .50 | .75 | .95 | avg | .05 | .25 | .50 | .75 | .95 |
| **DFM** | 0.152 | 0.039 | 0.107 | 0.150 | 0.193 | 0.271 | 0.036 | 0.003 | 0.014 | 0.029 | 0.053 | 0.093 |
| **Chow − Lin** | 0.390 | 0.062 | 0.235 | 0.385 | 0.526 | 0.739 | 0.120 | 0.007 | 0.041 | 0.093 | 0.171 | 0.334 |
| **Spline** | 0.160 | 0.077 | 0.118 | 0.154 | 0.194 | 0.265 | 0.036 | 0.003 | 0.014 | 0.029 | 0.051 | 0.091 |
| **K − filter** | 0.098 | 0.007 | 0.036 | 0.075 | 0.125 | 0.239 | 0.052 | 0.003 | 0.016 | 0.037 | 0.069 | 0.135 |
| **K − smoother** | 0.099 | 0.007 | 0.038 | 0.077 | 0.130 | 0.242 | 0.053 | 0.003 | 0.016 | 0.037 | 0.066 | 0.138 |

| | | MISSING OBSERVATIONS 40% | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\rho$ | | | | | | $\beta$ | | | | | |
| | avg | .05 | .25 | .50 | .75 | .95 | avg | .05 | .25 | .50 | .75 | .95 |
| **DFM** | 0.039 | 0.003 | 0.013 | 0.029 | 0.050 | 0.111 | 0.046 | 0.003 | 0.017 | 0.038 | 0.066 | 0.117 |
| **Chow − Lin** | 0.072 | 0.005 | 0.029 | 0.060 | 0.103 | 0.183 | 0.055 | 0.004 | 0.022 | 0.046 | 0.077 | 0.145 |
| **K − smoother** | 0.046 | 0.004 | 0.019 | 0.036 | 0.062 | 0.119 | 0.623 | 0.004 | 0.021 | 0.046 | 0.092 | 0.257 |

| | | MISSING OBSERVATIONS 5% | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\rho$ | | | | | | $\beta$ | | | | | |
| | avg | .05 | .25 | .50 | .75 | .95 | avg | .05 | .25 | .50 | .75 | .95 |
| **DFM** | 0.010 | 0.001 | 0.003 | 0.007 | 0.014 | 0.031 | 0.012 | 0.001 | 0.004 | 0.009 | 0.017 | 0.036 |
| **Chow − Lin** | 0.011 | 0.001 | 0.003 | 0.008 | 0.015 | 0.034 | 0.013 | 0.001 | 0.004 | 0.009 | 0.017 | 0.037 |
| **K − smoother** | 0.010 | 0.001 | 0.004 | 0.008 | 0.014 | 0.027 | 0.015 | 0.001 | 0.004 | 0.010 | 0.020 | 0.043 |

Note: The table reports the difference ($\rho$) between the first order autocorrelation coefficients for the actual and interpolated series, and the absolute value of the difference ($\beta$) of the estimated coefficient of $x_t$ in the regression $y_t = x_t + u_t$, with $u_t$ i.i.d. N(0,1), using actual and interpolated data for both $y_t$ and $x_t$. The DGP is as in (14).

**Table 8. Properties of interpolated data, empirical example**

| | ρ | | | | | | β | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **INFLATION** | | | | | | | | | | | |
| | AT | DE | ES | FI | FR | IT | AT | ES | FI | FR | IT |
| **DFM** | 0.161 | 0.086 | 0.003 | 0.065 | 0.027 | 0.001 | 0.089 | 0.047 | 0.021 | 0.019 | 0.002 |
| **Chow − Lin** | 0.116 | 0.101 | 0.004 | 0.068 | 0.023 | 0.006 | 0.035 | 0.016 | 0.016 | 0.025 | 0.001 |
| **Spline** | 0.183 | 0.183 | 0.30 | 0.083 | 0.037 | 0.009 | 0.101 | 0.049 | 0.010 | 0.015 | 0.017 |
| **K − filter** | 0.174 | 0.169 | 0.031 | 0.078 | 0.026 | 0.005 | 0.133 | 0.011 | 0.037 | 0.017 | 0.010 |
| **K − smoother** | 0.185 | 0.193 | 0.032 | 0.079 | 0.026 | 0.011 | 0.129 | 0.020 | 0.040 | 0.013 | 0.013 |

| | ρ | | | | | | β | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **MISSING OBSERVATIONS 20%** | | | | | | | | | | | |
| | AT | DE | ES | FI | FR | IT | AT | ES | FI | FR | IT |
| **DFM** | 0.040 | 0.061 | 0.028 | 0.031 | 0.012 | 0.002 | 0.020 | 0.056 | 0.053 | 0.079 | 0.019 |
| **Chow − Lin** | 0.004 | 0.029 | 0.015 | 0.026 | 0.011 | 0.004 | 0.008 | 0.004 | 0.007 | 0.017 | 0.014 |
| **K − smoother** | 0.002 | 0.018 | 0.026 | 0.028 | 0.000 | 0.008 | 0.160 | 0.117 | 0.134 | 0.237 | 0.128 |

| | ρ | | | | | | | β | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **REAL GDP GROWTH** | | | | | | | | | | | | | |
| | AT | DE | ES | FI | FR | IT | NL | AT | ES | FI | FR | IT | NL |
| **DFM** | 0.74 | 0.66 | 0.04 | 0.71 | 0.15 | 0.22 | 0.44 | 0.37 | 0.19 | 0.20 | 0.008 | 0.18 | 0.30 |
| **Chow − Lin** | 0.33 | 0.54 | 0.06 | 0.60 | 0.02 | 0.12 | 0.12 | 0.19 | 0.06 | 0.05 | 0.05 | 0.12 | 0.24 |
| **Spline** | 0.94 | 0.84 | 0.08 | 0.88 | 0.31 | 0.42 | 0.64 | 0.37 | 0.19 | 0.24 | 0.01 | 0.17 | 0.30 |
| **K − filter** | 0.84 | 0.77 | 0.07 | 0.84 | 0.26 | 0.31 | 0.57 | 0.37 | 0.19 | 0.24 | 0.009 | 0.18 | 0.31 |
| **K − smoother** | 0.85 | 0.77 | 0.07 | 0.87 | 0.29 | 0.32 | 0.57 | 0.37 | 0.19 | 0.24 | 0.03 | 0.18 | 0.31 |

| | ρ | | | | | | | β | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **MISSING OBSERVATIONS 20%** | | | | | | | | | | | | | |
| | AT | DE | ES | FI | FR | IT | NL | AT | ES | FI | FR | IT | NL |
| **DFM** | 0.18 | 0.06 | 0.03 | 0.04 | 0.07 | 0.06 | 0.19 | 0.14 | 0.03 | 0.01 | 0.03 | 0.02 | 0.03 |
| **Chow − Lin** | 0.24 | 0.05 | 0.03 | 0.06 | 0.04 | 0.07 | 0.05 | 0.15 | 0.08 | 0.02 | 0.02 | 0.25 | 0.00 |
| **K − smoother** | 0.18 | 0.06 | 0.03 | 0.04 | 0.09 | 0.06 | 0.20 | 0.13 | 0.04 | 0.006 | 0.05 | 0.23 | 0.03 |

Note: The table reports the difference ($\rho$) between the first order autocorrelation coefficients for the actual and interpolated series, and the absolute value of the difference ($\beta$) of the estimated coefficients in a regression of inflation or GDP growth for country $i$ on the same variable for country $j$, using actual and interpolated series.

AT: Austria, DE: Germany, ES: Spain, FI: Finland, FR: France, IT: Italy, NL: The Netherlands

# Data Appendix

Variables are denoted by three characters and countries by two.

CPI: Consumer Price Index, National Concept
MTD: Import Deflator
PCD: Private Consumption Deflator
PPI: Producers Price Index
XTD: Export Deflator
GCD: Government Consumption Deflator
ITD: Gross Fixed Capital Formation Deflator
YED: GDP Deflator

CAP: Capacity Utilizatiion
GDP: Real GDP
MTR: Real Imports
XTR: Real Exports
PCE: Private Consumption Expenditure
LTI: Long-term interest rate
STI: Short-term interest rate
LNN: Total Employment
UNN: Unemployment Rate
IIP: Industrial Production Total
AT: Austria
BE: Belgium
DE: Germany
ES: Spain
FI: Finland
FR: France
IE: Ireland
IT: Italy
NL: Netherlands
PT: Portugal

*List of variables in price dataset*

| | | |
|---|---|---|
| cpiat | pcdde | yedat |
| cpibe | pcdes | yedde |
| cpide | pcdfr | yedes |
| cpies | pcdfi | yedfi |
| cpifi | pcdit | yedfr |
| cpifr | ppiat | yedit |
| cpiie | ppide | gcdat |
| cpiit | ppies | gcdes |
| cpinl | ppifi | gcdfi |
| cpiptg | ppifr | gcdfr |
| mtdat | ppinl | gcdit |
| mtdde | xtdat | itdat |
| mtdes | xtdde | itdes |
| mtdfi | xtdes | itdfi |
| mtdfr | xtdfi | itdfr |
| mtdit | xtdfr | itdit |
| pcdat | xtdit | |

*List of variables in real dataset*

| | | |
|---|---|---|
| capde | pcees | lnnie |
| capes | pcefi | lnnit |
| capfr | pcefr | lnnnl |
| capit | pceit | lnnpt |
| capnl | pcenl | unrat |
| cappt | ltiat | unrbe |
| gdpat | ltibe | unrde |
| gdpde | ltide | unres |
| gdpes | ltifi | unrfi |
| gdpfi | ltifr | unrfr |
| gdpfr | ltiie | unrie |
| gdpit | ltiit | unrit |
| gdpnl | ltinl | unrnl |
| mtrat | stiat | unrpt |
| mtrde | stibe | iipatg |
| mtres | stide | iipbe |
| mtrfi | sties | iipde |
| mtrfr | stifi | iipes |
| mtrit | stifr | iipfi |
| mtrnl | stiie | iipfr |
| xtrat | stiit | iipie |
| xtrde | stinl | iipit |
| xtres | stipt | iipnl |
| xtrfi | lnnat | iippt |
| xtrfr | lnnbe | |
| xtrit | lnnde | |
| xtrnl | lnnes | |
| pceat | lnnfi | |
| pcede | lnnfr | |

# European Central Bank working paper series

For a complete list of Working Papers published by the ECB, please visit the ECB's website (http://www.ecb.int).

202 "Aggregate loans to the euro area private sector" by A. Calza, M. Manrique and J. Sousa, January 2003.

203 "Myopic loss aversion, disappointment aversion and the equity premium puzzle" by D. Fielding and L. Stracca, January 2003.

204 "Asymmetric dynamics in the correlations of global equity and bond returns" by L. Cappiello, R.F. Engle and K. Sheppard, January 2003.

205 "Real exchange rate in an inter-temporal n-country-model with incomplete markets" by B. Mercereau, January 2003.

206 "Empirical estimates of reaction functions for the euro area" by D. Gerdesmeier and B. Roffia, January 2003.

207 "A comprehensive model on the euro overnight rate" by F. R. Würtz, January 2003.

208 "Do demographic changes affect risk premiums? Evidence from international data" by A. Ang and A. Maddaloni, January 2003.

209 "A framework for collateral risk control determination" by D. Cossin, Z. Huang, D. Aunon-Nerin and F. González, January 2003.

210 "Anticipated Ramsey reforms and the uniform taxation principle: the role of international financial markets" by S. Schmitt-Grohé and M. Uribe, January 2003.

211 "Self-control and savings" by P. Michel and J.P. Vidal, January 2003.

212 "Modelling the implied probability of stock market movements" by E. Glatzer and M. Scheicher, January 2003.

213 "Aggregation and euro area Phillips curves" by S. Fabiani and J. Morgan, February 2003.

214 "On the selection of forecasting models" by A. Inoue and L. Kilian, February 2003.

215 "Budget institutions and fiscal performance in Central and Eastern European countries" by H. Gleich, February 2003.

216 "The admission of accession countries to an enlarged monetary union: a tentative assessment" by M. Ca'Zorzi and R. A. De Santis, February 2003.

217 "The role of product market regulations in the process of structural change" by J. Messina, March 2003.

218 "The zero-interest-rate bound and the role of the exchange rate for monetary policy in Japan" by G. Coenen and V. Wieland, March 2003.

219 "Extra-euro area manufacturing import prices and exchange rate pass-through" by B. Anderton, March 2003.

220 "The allocation of competencies in an international union: a positive analysis" by M. Ruta, April 2003.

221 "Estimating risk premia in money market rates" by A. Durré, S. Evjen and R. Pilegaard, April 2003.

222 "Inflation dynamics and subjective expectations in the United States" by K. Adam and M. Padula, April 2003.

223 "Optimal monetary policy with imperfect common knowledge" by K. Adam, April 2003.

224 "The rise of the yen vis-à-vis the ("synthetic") euro: is it supported by economic fundamentals?" by C. Osbat, R. Rüffer and B. Schnatz, April 2003.

225 "Productivity and the ("synthetic") euro-dollar exchange rate" by C. Osbat, F. Vijselaar and B. Schnatz, April 2003.

226 "The central banker as a risk manager: quantifying and forecasting inflation risks" by L. Kilian and S. Manganelli, April 2003.

227 "Monetary policy in a low pass-through environment" by T. Monacelli, April 2003.

228 "Monetary policy shocks – a nonfundamental look at the data" by M. Klaeffing, May 2003.

229 "How does the ECB target inflation?" by P. Surico, May 2003.

230 "The euro area financial system: structure, integration and policy initiatives" by P. Hartmann, A. Maddaloni and S. Manganelli, May 2003.

231 "Price stability and monetary policy effectiveness when nominal interest rates are bounded at zero" by G. Coenen, A. Orphanides and V. Wieland, May 2003.

232 "Describing the Fed's conduct with Taylor rules: is interest rate smoothing important?" by E. Castelnuovo, May 2003.

233 "The natural real rate of interest in the euro area" by N. Giammarioli and N. Valla, May 2003.

234 "Unemployment, hysteresis and transition" by M. León-Ledesma and P. McAdam, May 2003.

235 "Volatility of interest rates in the euro area: evidence from high frequency data" by N. Cassola and C. Morana, June 2003.

236 "Swiss monetary targeting 1974-1996: the role of internal policy analysis" by G. Rich, June 2003.

237 "Growth expectations, capital flows and international risk sharing" by O. Castrén, M. Miller and R. Stiegert, June 2003.

238 "The impact of monetary union on trade prices" by R. Anderton, R. E. Baldwin and D. Taglioni, June 2003.

239 "Temporary shocks and unavoidable transitions to a high-unemployment regime" by W. J. Denhaan, June 2003.

240 "Monetary policy transmission in the euro area: any changes after EMU?" by I. Angeloni and M. Ehrmann, July 2003.

241 Maintaining price stability under free-floating: a fearless way out of the corner?" by C. Detken and V. Gaspar, July 2003.

242 "Public sector efficiency: an international comparison" by A. Afonso, L. Schuknecht and V. Tanzi, July 2003.

243 "Pass-through of external shocks to euro area inflation" by E. Hahn, July 2003.

244 "How does the ECB allot liquidity in its weekly main refinancing operations? A look at the empirical evidence" by S. Ejerskov, C. Martin Moss and L. Stracca, July 2003.

245 "Money and payments: a modern perspective" by C. Holthausen and C. Monnet, July 2003.

246 "Public finances and long-term growth in Europe – evidence from a panel data analysis" by D. R. de Ávila Torrijos and R. Strauch, July 2003.

247 "Forecasting euro area inflation: does aggregating forecasts by HICP component improve forecast accuracy?" by K. Hubrich, August 2003.

248 "Exchange rates and fundamentals" by C. Engel and K. D. West, August 2003.

249 "Trade advantages and specialisation dynamics in acceding countries" by A. Zaghini, August 2003.

250 "Persistence, the transmission mechanism and robust monetary policy" by I. Angeloni, G. Coenen and F. Smets, August 2003.

251 "Consumption, habit persistence, imperfect information and the lifetime budget constraint" by A. Willman, August 2003.

252 "Interpolation and backdating with a large information set", by E. Angelini, J. Henry and M. Marcellino, August 2003.