3B2v8.06a/w (Dec 5 2003).51c
XML:ver:5.0.1
DYNCON : 1837
Prod.Type:FTP
pp.1–33(col.fig.:NIL)
ED:LathaN.K.
PAGN:Raj SCAN:Nil

ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT

ELSEVIER

Journal of Economic Dynamics & Control ▮ (▮▮▮▮) ▮▮▮–▮▮▮

JOURNAL OF Economic Dynamics & Control

www.elsevier.com/locate/jedc

Interpolation and backdating with a large information set

Elena Angelini[a], Jérôme Henry[a], Massimiliano Marcellino[b,*]

[a]DG-Research, ECB

[b]IEP - Bocconi University, IGIER and CEPR, Via Salasco 5, 20136, Milan, Italy

Received 22 September 2003; accepted 5 July 2005

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 easily 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 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 rather 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, but they are smaller with multivariate approaches, including factor-based ones.

JEL classification: C32; C43; C82

Keywords: Interpolation; Factor model; Kalman filter; Spline

*Corresponding author. Tel.: +39 2 5836 3327; fax: +39 2 5836 3302.
E-mail address: massimiliano.marcellino@unibocconi.it (M. Marcellino).

0165-1889/$ - see front matter © 2005 Published by Elsevier B.V.
doi:10.1016/j.jedc.2005.07.010

1. Introduction

This paper proposes a factor-based procedure to interpolate economic data (e.g. obtain monthly values from quarterly values), backdate time series (i.e. estimate observations before the starting date of the available sample) and deal with missing observations and outliers, using large information sets.

These issues received considerable attention in the literature. A first, simple, approach to recovering the disaggregated values is to compute 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 minimize a loss function under a compatibility constraint with aggregated data, see e.g. Boot et al. (1967), Cohen et al. (1971), and Stram and Wei (1986). A further constraint can be added, involving 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 interpolation 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).

The common idea underlying all of these methods, reviewed e.g. in Marcellino (1998), is to project the series of interest, which contains missing observations, either on its own available data or on a set of series covering the whole sample. The traditional multivariate procedures are usually applied using a limited number of regressors because of the typical curse of dimensionality problem. For this reason, a pre-selection of the regressors has to be made prior to applying such methods.

An alternative procedure to overcome the curse of dimensionality problem is to model the large amount of available information with a factor model, where all variables are driven by a limited number of common factors, and use the estimated factors as regressors. Traditional factor models are not suited for economic applications, 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 over time, see Stock and Watson (2002a,b) and Forni et al. (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. The estimated factors, using a principal component-based estimation procedure, are particularly useful for forecasting, which can be considered as a problem of missing observations at the end of the series, see e.g. Stock and Watson (2002a) for the US, Marcellino et al. (2003) and Angelini et al. (2001) for the Euro area, and Artis et al. (2004) for the UK.

In this paper we apply the idea of summarizing large information sets with a few estimated factors to the problems of data interpolation and series backdating. We compare the factor-based approach 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, develop the factor-based estimators, and compare them with competing methods from a theoretical point of view. While the framework is very simple and based on linear projection theory, it provides a convenient environment to derive most of the common disaggregation methods and compare them with the factor-based approach. In Section 3 we evaluate the relative merits of the disaggregation methods by means of a comprehensive set of simulation experiments. In Section 4 we apply the methods to some macroeconomic variables for European countries and for the euro area, to evaluate their practical performance. In Section 5 we analyze the consequences of using the interpolated/backdated data in subsequent econometric analyses, using both simulation experiments and the European macroeconomic data. Finally, in Section 6 we summarize the main findings of the paper and conclude.

2. Factor-based interpolation and backdating

In this section we introduce the framework to analyze the interpolation/backdating problem, review some of the existing methods, develop the new factor-based approach, and compare the competing procedures.

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 variables. Precise conditions on the factors, F_t, and the idiosyncratic errors, e_t, can be found in Stock and Watson (2002a,b).

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

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

Yet, not all values of y_t^{o} can be observed. In particular, observed values can be thought of as realizations of the process $y = \{y_\tau\}_{\tau=1}^{\infty} = \{\omega(L) y_{kt}^{\mathrm{o}}\}_{t=1}^{\infty}$, where τ 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 $\omega(L) = \omega_0 + \omega_1 L + \cdots + \omega_{k-1} L^{k-1}$ characterizes the aggregation scheme. For example, $\omega(L) = 1 + L + \cdots + L^{k-1}$ 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_τ in $\underset{nT\times 1}{\mathbf{X}}$, $\underset{T\times 1}{\mathbf{Y}^{\mathrm{o}}}$ and $\underset{s\times 1}{\mathbf{Y}}$, where s is the number of aggregate observations, and construct the aggregator matrix $\overline{\mathbf{W}}$ with

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

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

then $\mathbf{Z} = \overline{\mathbf{W}}\mathbf{Z}^{\mathrm{o}}$, where $\mathbf{Z}^{\mathrm{o}} = (\mathbf{Y}^{\mathrm{o}\prime} : \mathbf{X}')'$ 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. Similarly, the structure of $\mathbf{W}$ can be easily modified to deal with missing observations at the beginning of the sample, or elsewhere. In particular, since the processes under analysis are weakly stationary, the results for backdating are also applicable to the case of forecasting, i.e. missing observations on the y variable at the end of the sample, assuming that the corresponding values of the X variables are known.

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

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

This formulation is very general. 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

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

Using the Choleski decomposition $\mathbf{Q} = \mathbf{PP}'$ and defining $\mathbf{R}^{\mathrm{o}} = \mathbf{Z}^{\mathrm{o}}\mathbf{P}^{-1}$, $\mathbf{R} = \mathbf{ZP}^{-1}$, $\widetilde{\mathbf{R}} = \widetilde{\mathbf{Z}}\mathbf{P}^{-1}$, (4) can be written as (3), after substituting $\mathbf{Z}$ with $\mathbf{R}$.

A large class of interpolation methods can be obtained by solving the problem in (3), possibly with some additional conditions or constraints. Therefore, while the solution of the optimization problem in (3) is simple, we think it is worth analyzing it in some details. It also facilitates the theoretical comparison across various methods, including those using factors.

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

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

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

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

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

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

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

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

$$\widehat{\mathbf{Z}} = \begin{pmatrix} \widehat{\mathbf{Y}} \\ \widehat{\mathbf{X}} \end{pmatrix} = \begin{pmatrix} \underset{T\times s}{\boldsymbol{\alpha}} & \underset{T\times nT}{\boldsymbol{\beta}} \\ \underset{nT\times s}{\boldsymbol{\gamma}} & \underset{nT\times nT}{\boldsymbol{\delta}} \end{pmatrix} \begin{pmatrix} \mathbf{Y} \\ \mathbf{X} \end{pmatrix}, \tag{5}$$

where

$$\boldsymbol{\alpha} = (\mathbf{V}_{\mathbf{Y}^{\circ}} - \mathbf{C}_{\mathbf{Y}^{\circ}\mathbf{X}}\mathbf{V}_{\mathbf{X}}^{-1}\mathbf{C}_{\mathbf{X}\mathbf{Y}^{\circ}})\mathbf{W}'[\mathbf{W}(\mathbf{V}_{\mathbf{Y}^{\circ}} - \mathbf{C}_{\mathbf{Y}^{\circ}\mathbf{X}}\mathbf{V}_{\mathbf{X}}^{-1}\mathbf{C}_{\mathbf{X}\mathbf{Y}^{\circ}})\mathbf{W}']^{-1},$$

$$\boldsymbol{\beta} = [\mathbf{I} - \mathbf{V}_{\mathbf{Y}^{\circ}}\mathbf{W}'(\mathbf{W}\mathbf{V}_{\mathbf{Y}^{\circ}}\mathbf{W}')^{-1}\mathbf{W}]\mathbf{C}_{\mathbf{Y}^{\circ}\mathbf{X}}[\mathbf{V}_{\mathbf{X}} - \mathbf{C}_{\mathbf{X}\mathbf{Y}^{\circ}}\mathbf{W}'(\mathbf{W}\mathbf{V}_{\mathbf{Y}^{\circ}}\mathbf{W}')^{-1}\mathbf{W}\mathbf{C}_{\mathbf{Y}^{\circ}\mathbf{X}}]^{-1},$$

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

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

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 $\mathbf{Y}^{\circ}$ on $\mathbf{Y}$ and $\mathbf{X}$. In an obvious notation, we have

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

$$\boldsymbol{\beta} = \mathbf{V}_{\mathbf{Y}^{\circ}|\mathbf{Y}}\mathbf{W}'\mathbf{V}_{\mathbf{X}|\mathbf{Y}}^{-1}.$$

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, ΛF_t and e_t, respectively. Stacking F_t and e_t into $\mathbf{F}$ and $\mathbf{e}$, we have,

[1]The proofs of all the results in this section are rather simple and can be found in the working paper version of this article, see Angelini et al. (2003).

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

$$\widehat{\mathbf{Z}}_F = \begin{pmatrix} \boldsymbol{\alpha}_F & \boldsymbol{\beta}_F \\ \boldsymbol{\gamma}_F & \boldsymbol{\delta}_F \end{pmatrix} \begin{pmatrix} \mathbf{Y} \\ \mathbf{F} \end{pmatrix}, \tag{6}$$

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

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

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

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

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

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* $n = p$.

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* $\text{cov}(\mathbf{Y}^o, \mathbf{X}|\mathbf{Y}) = 0$, *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*,

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

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

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

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

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

$$E(\mathbf{Y}^o - \widehat{\mathbf{Y}}_U)(\mathbf{Y}^o - \widehat{\mathbf{Y}}_U)' = \mathbf{V}_{\mathbf{Y}^o} - \mathbf{V}_{\mathbf{Y}^o}\mathbf{W}'\mathbf{V}_{\mathbf{Y}}^{-1}\mathbf{W}\mathbf{V}_{\mathbf{Y}^o}.$$

$\widehat{\mathbf{Z}}_U$ 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*

$$\min \operatorname{tr}(E(\mathbf{Z}^{\mathrm{o}} - \widetilde{\mathbf{Z}})(\mathbf{Z}^{\mathrm{o}} - \widetilde{\mathbf{Z}})' | \mathbf{C}_{\mathbf{Y}^{\mathrm{o}}\mathbf{X}}\mathbf{V}_{\mathbf{X}}\mathbf{X}) \quad s.t. \quad \mathbf{Z} = \overline{\mathbf{W}}\mathbf{Z}^{\mathrm{o}} \tag{8}$$

is

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

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

$$\boldsymbol{\alpha}_{\mathrm{C}} = (\mathbf{V}_{\mathbf{Y}^{\mathrm{o}}} - \mathbf{C}_{\mathbf{Y}^{\mathrm{o}}\mathbf{X}}\mathbf{V}_{\mathbf{X}}^{-1}\mathbf{C}_{\mathbf{X}\mathbf{Y}^{\mathrm{o}}})\mathbf{W}'[\mathbf{W}(\mathbf{V}_{\mathbf{Y}^{\mathrm{o}}} - \mathbf{C}_{\mathbf{Y}^{\mathrm{o}}\mathbf{X}}\mathbf{V}_{\mathbf{X}}^{-1}\mathbf{C}_{\mathbf{X}\mathbf{Y}^{\mathrm{o}}})\mathbf{W}']^{-1},$$

$$\boldsymbol{\beta}_{\mathrm{C}} = [\mathbf{I} - \boldsymbol{\alpha}_{\mathrm{C}}]\mathbf{C}_{\mathbf{Y}^{\mathrm{o}}\mathbf{X}}\mathbf{V}_{\mathbf{X}},$$

$$\boldsymbol{\gamma}_{\mathrm{C}} = \mathbf{0},$$

$$\boldsymbol{\delta}_{\mathrm{C}} = \mathbf{I}.$$

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

$$\widehat{\mathbf{Z}}_{\mathrm{C}} = \widehat{\mathbf{Z}}.$$

We call $\widehat{\mathbf{Z}}_{\mathrm{C}}$ the *conditional estimator*. Notice that $\widehat{\mathbf{Y}}_{\mathrm{C}}$ 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 $\widehat{\mathbf{Y}}_{\mathrm{C}}$ in two steps. In the first step $\mathbf{Y}^{\mathrm{o}}$ is projected on $\mathbf{X}$. In the second step, the residuals from the first step are projected on their aggregated counterpart. If $\mathbf{Y}$ and $\mathbf{X}$ are uncorrelated, this procedure is equivalent to projecting $\mathbf{Y}^{\mathrm{o}}$ 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, $\mathbf{Y}_p^{\mathrm{o}}$, but it does not satisfy the aggregation constraint $\mathbf{Y} = \mathbf{W}\mathbf{Y}_p^{\mathrm{o}}$. In this case the problem becomes

$$\min_{\widetilde{\mathbf{Z}}} \operatorname{tr}(E(\mathbf{Z}^{\mathrm{o}} - \widetilde{\mathbf{Z}})(\mathbf{Z}^{\mathrm{o}} - \widetilde{\mathbf{Z}})' | \mathbf{Y}_p^{\mathrm{o}}) \quad \text{s.t.} \quad \mathbf{Z} = \overline{\mathbf{W}}\mathbf{Z}^{\mathrm{o}}, \tag{10}$$

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

$$\widehat{\mathbf{Y}}_{\mathrm{P}} = \mathbf{Y}_p^{\mathrm{o}} + \mathbf{V}_{\mathbf{Y}^{\mathrm{o}}}\mathbf{W}'\mathbf{V}_{\mathbf{Y}}^{-1}(\mathbf{Y} - \mathbf{W}\mathbf{Y}_p^{\mathrm{o}}). \tag{11}$$

We refer to $\widehat{\mathbf{Y}}_{\mathrm{P}}$ as to the *preliminary estimator*. $\widehat{\mathbf{Y}}_{\mathrm{P}}$ boils down to the conditional estimator $\widehat{\mathbf{Y}}_{\mathrm{C}}$ when $\mathbf{Y}_p^{\mathrm{o}} = \mathbf{C}_{\mathbf{Y}^{\mathrm{o}}\mathbf{X}}\mathbf{V}_{\mathbf{X}}\mathbf{X}$. Chow and Lin's (1971) estimator belongs to this class. In their case $\mathbf{Y}_p^{\mathrm{o}} = \widehat{\gamma}_{GLS}\mathbf{X}$, 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 $\widehat{\mathbf{Y}}_{\mathrm{F}}$, $\widehat{\mathbf{Y}}_{\mathrm{U}}$, and $\widehat{\mathbf{Y}}_{\mathrm{C}}$ 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,

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

$$\underset{(n+1)T\times 1}{\boldsymbol{\varepsilon}} = \begin{pmatrix} \underset{T\times 1}{\mathbf{0}} \\ \underset{nT\times 1}{\mathbf{e}} \end{pmatrix},$$

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

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

$$\boldsymbol{\Sigma}_{\text{F}} - \boldsymbol{\Sigma} = (\mathbf{aA} - \mathbf{a}_{\text{F}})\mathbf{V}_{\mathbf{Z}_{\text{F}}}(\mathbf{aA} - \mathbf{a}_{\text{F}})' + (\mathbf{aA} - \mathbf{a}_{\text{F}})\mathbf{C}_{\mathbf{Z}_{\text{F}}\boldsymbol{\varepsilon}} + \mathbf{C}_{\boldsymbol{\varepsilon}}Z_{\text{F}}(\mathbf{aA} - \mathbf{a}_{\text{F}})' + \mathbf{V}_{\boldsymbol{\varepsilon}},$$

$$\boldsymbol{\Sigma}_{\text{U}} - \boldsymbol{\Sigma} = (\mathbf{a} - \mathbf{a}_{\text{U}})\mathbf{V}_{\mathbf{Z}^{\text{o}}}(\mathbf{a} - \mathbf{a}_{\text{U}})',$$

$$\boldsymbol{\Sigma}_{\text{C}} - \boldsymbol{\Sigma} = (\mathbf{a} - \mathbf{a}_{\text{C}})\mathbf{V}_{\mathbf{Z}^{\text{o}}}(\mathbf{a} - \mathbf{a}_{\text{C}})'.$$

Table 1 summarizes the alternative interpolation (and/or backdating) methods.

Table 1
Alternative estimators

Joint : $\widehat{\mathbf{Y}} = \boldsymbol{\alpha}\mathbf{Y} + \boldsymbol{\beta}\mathbf{X}$
$\boldsymbol{\alpha} = (\mathbf{V}_{\mathbf{Y}^{\text{o}}} - \mathbf{C}_{\mathbf{Y}^{\text{o}}\mathbf{X}}\mathbf{V}_{\mathbf{X}}^{-1}\mathbf{C}_{\mathbf{X}\mathbf{Y}^{\text{o}}})\mathbf{W}'[\mathbf{W}(\mathbf{V}_{\mathbf{Y}^{\text{o}}} - \mathbf{C}_{\mathbf{Y}^{\text{o}}\mathbf{X}}\mathbf{V}_{\mathbf{X}}^{-1}\mathbf{C}_{\mathbf{X}\mathbf{Y}^{\text{o}}})\mathbf{W}']^{-1}$
$\boldsymbol{\beta} = [\mathbf{I} - \mathbf{V}_{\mathbf{Y}^{\text{o}}}\mathbf{W}'(\mathbf{W}\mathbf{V}_{\mathbf{Y}^{\text{o}}}\mathbf{W}')^{-1}\mathbf{W}]\mathbf{C}_{\mathbf{Y}^{\text{o}}\mathbf{X}}[\mathbf{V}_{\mathbf{X}} - \mathbf{C}_{\mathbf{X}\mathbf{Y}^{\text{o}}}\mathbf{W}'(\mathbf{W}\mathbf{V}_{\mathbf{Y}^{\text{o}}}\mathbf{W}')^{-1}\mathbf{W}\mathbf{C}_{\mathbf{Y}^{\text{o}}\mathbf{X}}]^{-1}$
Factor : $\widehat{\mathbf{Y}}_{\text{F}} = \boldsymbol{\alpha}_{\text{F}}\mathbf{Y} + \boldsymbol{\beta}_{\text{F}}\mathbf{F}$
$\boldsymbol{\alpha}_{\text{F}} = (\mathbf{V}_{\mathbf{Y}^{\text{o}}} - \mathbf{C}_{Y^{\text{o}}\text{F}}\mathbf{V}_{\text{F}}^{-1}\mathbf{C}_{\text{F}Y^{\text{o}}})\mathbf{W}'[\mathbf{W}(\mathbf{V}_{\mathbf{Y}^{\text{o}}} - \mathbf{C}_{Y^{\text{o}}\text{F}}\mathbf{V}_{\text{F}}^{-1}\mathbf{C}_{\text{F}Y^{\text{o}}})\mathbf{W}']^{-1}$
$\boldsymbol{\beta}_{\text{F}} = [\mathbf{I} - \mathbf{V}_{\mathbf{Y}^{\text{o}}}\mathbf{W}'(\mathbf{W}\mathbf{V}_{Y^{\text{o}}}\mathbf{W}')^{-1}\mathbf{W}]\mathbf{C}_{Y^{\text{o}}\text{F}}[\mathbf{V}_{\text{F}} - \mathbf{C}_{\text{F}Y^{\text{o}}}\mathbf{W}'(\mathbf{W}\mathbf{V}_{Y^{\text{o}}}\mathbf{W}')^{-1}\mathbf{W}\mathbf{C}_{Y^{\text{o}}\text{F}}]^{-1}$
Univariate : $\widehat{\mathbf{Y}}_{\text{U}} = \boldsymbol{\alpha}_{\text{U}}\mathbf{Y}$
$\boldsymbol{\alpha}_{\text{U}} = \mathbf{V}_{\mathbf{Y}^{\text{o}}}\mathbf{W}'\mathbf{V}_{\mathbf{Y}}^{-1}$
Conditional : $\widehat{\mathbf{Y}}_{\text{C}} = \boldsymbol{\alpha}_{\text{C}}\mathbf{Y} + \boldsymbol{\beta}_{\text{C}}\mathbf{X}$
$\boldsymbol{\alpha}_{\text{C}} = (\mathbf{V}_{\mathbf{Y}^{\text{o}}} - \mathbf{C}_{\mathbf{Y}^{\text{o}}\mathbf{X}}\mathbf{V}_{\mathbf{X}}^{-1}\mathbf{C}_{\mathbf{X}\mathbf{Y}^{\text{o}}})\mathbf{W}'[\mathbf{W}(\mathbf{V}_{\mathbf{Y}^{\text{o}}} - \mathbf{C}_{\mathbf{Y}^{\text{o}}\mathbf{X}}\mathbf{V}_{\mathbf{X}}^{-1}\mathbf{C}_{\mathbf{X}\mathbf{Y}^{\text{o}}})\mathbf{W}']^{-1}$
$\boldsymbol{\beta}_{\text{C}} = [\mathbf{I} - \boldsymbol{\alpha}_{\text{C}}]\mathbf{C}_{\mathbf{Y}^{\text{o}}\mathbf{X}}\mathbf{V}_{\mathbf{X}}$
Preliminary : $\widehat{\mathbf{Y}}_{\text{P}} = \mathbf{Y}_p^{\text{o}} + \boldsymbol{\alpha}_{\text{U}}(\mathbf{Y} - \mathbf{W}\mathbf{Y}_p^{\text{o}})$
$\boldsymbol{\alpha}_{\text{U}} = \mathbf{V}_{\mathbf{Y}^{\text{o}}}\mathbf{W}'\mathbf{V}_{\mathbf{Y}}^{-1}$

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

3. 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 third subsection we discuss the results. In the final subsection we comment upon the outcome of a set of sensitivity analyses.

3.1. 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 detail 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, the estimation of the high-order lags of the autocovariance function is highly imprecise in finite samples. Moreover, several elements in these matrices are likely to be 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 disaggregate 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. This reflects common empirical practices, where, typically, variables involved in Chow–Lin procedures are selected on the basis of e.g. their economic relevance with respect to the estimated variable.

Next, we consider two types of factor-based estimators. The first 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 algorithm developed by Stock and Watson (2002a,b). Basically, the procedure starts by the disaggregated variable obtained by the first factor method to the balanced panel, factors are re-extracted, the Chow and Lin (1971) procedure is then 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 (2002a,b) 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) and Bai and Ng (2004).[2]

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

3.2. The experimental design

We consider three different generating mechanisms (DGP) for the variables:

$$X_t = \Lambda F_t + e_t,$$

$$y_t^{\mathrm{o}} = \beta' F_t + \varepsilon_t, \tag{12}$$

$$X_t = \Lambda F_t + e_t,$$

$$y_t^{\mathrm{o}} = \beta' Z_t + \varepsilon_t, \tag{13}$$

and

$$\begin{aligned} X_t &= Q X_{t-1} + u_t, \\ y_t^{\mathrm{o}} &= \gamma y_{t-1}^{\mathrm{o}} + v_t, \end{aligned} \qquad \begin{pmatrix} u_t \\ v_t \end{pmatrix} = P \begin{pmatrix} e_t \\ \varepsilon_t \end{pmatrix}. \tag{14}$$

[2] Notice that it could be possible to pre-select the variables prior to factor estimation, along the lines of the procedure suggested for the Chow–Lin approach. Boivin and Ng (2004) showed that this can yield some gains in a forecasting context.

In the first specification (12) both the X and the y variables are generated by a factor model. 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 Λ and β are independent draws from a uniform distribution over the interval [0, 1]. In the second specification (13) $Z_t = (x_{1t}, x_{2t}, x_{3t})'$, so that y depends on some of the variables in X rather than on the factors. In the third specification (14) y and each of the variables in X are AR(1) processes, each with root equal to 0.8 (Q is a diagonal matrix).

In the first two DGPs e_t and ε_t are i.i.d. $N(0,1)$ errors, uncorrelated across themselves, while in the third DGP the errors are correlated across variables (the elements of P are independent draws from a uniform distribution), but not over time. In all cases X_t contains 50 variables while y_t^{o} is univariate, and the sample size is set equal to 100.[3]

When the DGP is (12) 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 the DGP is (13) the Chow–Lin method is expected to generate the lowest loss function, but the factor-based interpolation approach could also perform well when the factors have a high explanatory power for the Z variables, since the model for y_t^{o} in (13) can be written as

$$y_t^{\mathrm{o}} = \beta' S \Lambda F_t + \beta' S e_t, + \varepsilon_t, \tag{15}$$

where S is the selection matrix

$$S = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \end{pmatrix}.$$

When data are generated according to (14) the univariate estimators should be ranked first, but the multivariate methods could also perform well due to the correlation in the error terms. Hence, though more complicated DGPs could be used, those in (12)–(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 out of 100. 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

[3]The simulation experiments in Banerjee et al. (2005) indicate that increasing the cross-section dimension or allowing for cross-correlation in the idiosyncratic errors do not affect the forecasting performance of the factor model. Therefore, we can expect these two features of the experimental design not to matter also in our context.

the absolute and mean square disaggregation error, which provides additional information on the performance of the estimators.

3.3. 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.

Table 2
Disaggregation error, DGM is DFM

	MSE						MAE					
	avg	.05	.25	.50	.75	.95	avg	.05	.25	.50	.75	.95
Stock												
DFM	0.304	0.132	0.198	0.269	0.370	0.608	0.373	0.252	0.308	0.360	0.424	0.538
Chow–Lin	0.380	0.174	0.256	0.342	0.459	0.727	0.417	0.288	0.348	0.405	0.470	0.596
Spline	1.166	0.727	0.961	1.154	1.353	1.634	0.735	0.586	0.675	0.739	0.797	0.876
K-filter	0.859	0.603	0.743	0.832	0.944	1.182	0.639	0.535	0.594	0.635	0.675	0.760
K-smoother	1.219	0.593	0.736	0.833	0.954	1.230	0.644	0.529	0.593	0.634	0.677	0.768
Fraction of cases where balanced panel works better than non-balanced panel: 1.000												
Flow												
DFM	0.290	0.129	0.192	0.263	0.355	0.547	0.420	0.285	0.351	0.409	0.478	0.593
Chow–Lin	0.364	0.171	0.250	0.332	0.442	0.662	0.472	0.328	0.400	0.461	0.533	0.650
Spline	0.624	0.394	0.536	0.631	0.724	0.827	0.629	0.501	0.583	0.636	0.683	0.736
K-filter	0.644	0.432	0.559	0.648	0.732	0.828	0.640	0.526	0.594	0.644	0.687	0.736
K-smoother	0.641	0.424	0.554	0.646	0.730	0.829	0.638	0.521	0.592	0.645	0.686	0.737
Fraction of cases where balanced panel works better than non-balanced panel: 0.914												
Missing observations 40%												
DFM	0.153	0.063	0.099	0.135	0.186	0.307	0.193	0.128	0.159	0.186	0.220	0.282
Chow–Lin	0.177	0.077	0.118	0.159	0.217	0.339	0.208	0.138	0.174	0.203	0.236	0.297
K-smoother	1.351	0.271	0.368	0.443	0.536	0.926	0.382	0.263	0.308	0.339	0.378	0.504
Fraction of cases where balanced panel works better than non-balanced panel: 1.000												
Missing observations 5%												
DFM	0.018	0.003	0.009	0.015	0.024	0.045	0.024	0.011	0.017	0.022	0.029	0.041
Chow–Lin	0.021	0.004	0.010	0.016	0.027	0.050	0.025	0.011	0.018	0.024	0.031	0.043
K-smoother	0.047	0.010	0.023	0.039	0.063	0.113	0.039	0.019	0.029	0.037	0.048	0.065
Fraction of cases where balanced panel works better than non-balanced panel: 1.000												

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 (12), for different disaggregation methods, types of variables and of missing observations.

Table 3
Disaggregation error, DGM is Chow–Lin

	MSE						MAE					
	avg	.05	.25	.50	.75	.95	avg	.05	.25	.50	.75	.95
Stock												
DFM	0.284	0.117	0.179	0.251	0.359	0.572	0.359	0.235	0.293	0.347	0.417	0.528
Chow–Lin	0.253	0.073	0.139	0.210	0.323	0.576	0.333	0.188	0.258	0.319	0.395	0.528
Spline	1.182	0.766	0.991	1.175	1.352	1.631	0.740	0.597	0.683	0.745	0.799	0.875
K-filter	0.865	0.626	0.748	0.846	0.954	1.161	0.642	0.543	0.598	0.640	0.680	0.751
K-smoother	1.048	0.610	0.751	0.848	0.958	1.207	0.678	0.538	0.599	0.639	0.682	0.763
Fraction of cases where balanced panel works better than non-balanced panel: 0.880												
Flow												
DFM	0.271	0.111	0.178	0.243	0.333	0.537	0.406	0.268	0.336	0.393	0.462	0.590
Chow–Lin	0.251	0.079	0.146	0.218	0.319	0.547	0.385	0.225	0.306	0.371	0.451	0.591
Spline	0.623	0.420	0.540	0.628	0.711	0.812	0.628	0.514	0.586	0.632	0.675	0.728
K-filter	1.856	0.434	0.564	0.643	0.723	0.827	0.660	0.526	0.599	0.644	0.684	0.735
K-smoother	1.017	0.431	0.559	0.640	0.722	0.828	0.646	0.525	0.596	0.642	0.683	0.735
Fraction of cases where balanced panel works better than non-balanced panel: 0.918												
Missing observations 40%												
DFM	0.141	0.055	0.087	0.124	0.175	0.288	0.185	0.119	0.149	0.179	0.214	0.273
Chow–Lin	0.108	0.032	0.058	0.089	0.136	0.248	0.159	0.090	0.122	0.152	0.187	0.255
K-smoother	1.307	0.276	0.371	0.445	0.536	0.951	0.389	0.265	0.310	0.340	0.378	0.523
Fraction of cases where balanced panel works better than non-balanced panel: 0.634												
Missing observations 5%												
DFM	0.018	0.003	0.007	0.014	0.024	0.047	0.023	0.010	0.016	0.022	0.029	0.042
Chow–Lin	0.013	0.002	0.005	0.009	0.017	0.038	0.019	0.008	0.013	0.018	0.024	0.037
K-smoother	0.049	0.010	0.024	0.041	0.064	0.113	0.039	0.018	0.029	0.038	0.048	0.064
Fraction of cases where balanced panel works better than non-balanced panel: 0.480												

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.

A first finding, robust across experiments, 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 second best is Chow–Lin, with an increase of about 20% in the loss function. The univariate methods do not perform

Table 4
Disaggregation error, DGM is AR(1)

	MSE						MAE					
	avg	.05	.25	.50	.75	.95	avg	.05	.25	.50	.75	.95
Stock												
DFM	0.625	0.490	0.559	0.608	0.671	0.806	0.550	0.479	0.519	0.545	0.576	0.634
Chow–Lin	0.559	0.340	0.447	0.540	0.650	0.836	0.513	0.401	0.462	0.510	0.560	0.642
Spline	0.549	0.291	0.414	0.521	0.656	0.877	0.491	0.368	0.435	0.488	0.543	0.624
K-filter	0.736	0.384	0.545	0.688	0.851	1.187	0.581	0.426	0.510	0.572	0.639	0.759
K-smoother	1.791	0.305	0.483	0.639	0.824	1.221	0.572	0.379	0.480	0.553	0.627	0.760
Fraction of cases where balanced panel works better than non-balanced panel: 0.925												
Flow												
DFM	0.988	0.372	0.616	0.844	1.188	1.982	0.772	0.490	0.629	0.739	0.880	1.135
Chow–Lin	1.198	0.493	0.749	0.999	1.413	2.493	0.850	0.558	0.692	0.804	0.953	1.267
Spline	0.239	0.129	0.181	0.229	0.286	0.380	0.380	0.283	0.336	0.377	0.421	0.486
K-filter	0.385	0.180	0.275	0.364	0.464	0.636	0.477	0.336	0.411	0.473	0.532	0.627
K-smoother	0.370	0.169	0.257	0.343	0.452	0.635	0.484	0.324	0.398	0.460	0.524	0.626
Fraction of cases where balanced panel works better than non-balanced panel: 0.950												
Missing observations 40%												
DFM	0.296	0.164	0.234	0.288	0.352	0.454	0.278	0.204	0.246	0.277	0.307	0.353
Chow–Lin	0.288	0.157	0.221	0.278	0.344	0.450	0.272	0.200	0.240	0.270	0.302	0.350
K-smoother	1.737	0.229	0.357	0.459	0.599	0.909	0.371	0.240	0.304	0.349	0.401	0.516
Fraction of cases where balanced panel works better than non-balanced panel: 0.778												
Missing observations 5%												
DFM	0.038	0.005	0.014	0.026	0.049	0.112	0.035	0.013	0.022	0.031	0.044	0.070
Chow–Lin	0.035	0.006	0.015	0.026	0.045	0.094	0.033	0.014	0.023	0.030	0.041	0.062
K-smoother	0.042	0.006	0.014	0.028	0.053	0.127	0.036	0.014	0.022	0.032	0.045	0.073
Fraction of cases where balanced panel works better than non-balanced panel: 0.746												

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.

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 on the y variable are generated by the Chow–Lin specification, Table 3 indicates that the Chow–Lin interpolation/backdating method becomes the best, the factor-based approach is the second best with a loss of about 20%, and the bad results for the univariate methods are confirmed.

When the data are generated by the SUR-type DGP in (14), in turn, univariate methods would be expected to provide the best estimates, but the results in Table 4 show that this is not a clear-cut case. For interpolation of flow variables, the Spline method is the best, with the Kalman filter and smoother performing about 30–40% worse, and the multivariate methods are the worst. For stock variables, the Spline method remains the best but the performance of the multivariate approaches improves substantially. This is even more evident in the case of backdating, when both Chow–Lin and the factor- based approaches outperform the Kalman smoother, in particular when the number of missing observations is high.

In summary, the multivariate methods appear to perform quite well in the simulation experiments, even when they are based on a misspecified model. The ranking of the traditional Chow–Lin method (supplemented with a careful variable selection) and of the new factor-based approach is not clear cut. Therefore, we consider additional simulation experiments to shed light on this issue.

3.4. Sensitivity analysis

The first additional experiment we consider deals with the misspecification of the number of factors. We use the specification in (12) to generate the data, but there are 10 factors in the DGP while only five of them are used in the factor-based interpolation procedure. The results, reported in Table 5, indicate that the factor method still substantially outperforms the univariate approaches, but Chow–Lin remains a valid alternative.

In the second additional experiment we use the DGP in (13) but now y depends on all the variables in X, i.e. $Z_t = X_t$ in (13). This is the case for which the factor-based method should be the most appropriate, since the y variable depends on 50 regressors and there are not enough observations to run a Chow–Lin procedure with all the potentially relevant variables included. Therefore, we can either summarize the 50 variables by means of (three) estimated factors or select those (five variables) that present the highest correlation with y. The figures in Table 6 are impressive. Now for all types of interpolation or backdating the losses from the Chow–Lin method are 10 times larger than those from the factor approach.

The third set of additional experiments increases the variability of the idiosyncratic error of the factor model in (12) and (13), specifically the idiosyncratic error is $2e_t$ rather than e_t. This complicates the estimation of the factors and reduces their explanatory power, so that the performance of factor-based interpolation could be expected to deteriorate. On the other hand, since the variability of the error term in the y equation, ε_t in (12) and (13), remains the same, while the variance of each of the variables in X increases, the Chow–Lin method could perform even better than in the standard case with the DGP in (13). The results reported in Table 7 for the DGP (12) and 8 for (13) support our expectations. From Table 7, the factor-based method deteriorates by about 20% with respect to the values reported in Table 2, while the losses increase to 30–40% in Table 8 with respect to Table 3. The Chow–Lin method suffers from the same losses as the factor method with the modified factor DGP (Table 7), while it improves substantially in Table 8 with respect to Table 3, as expected. In both cases, the multivariate methods remain by far superior to the univariate procedures.

DYNCON : 1837

ARTICLE IN PRESS

Table 5
Disaggregation error, DGM is DFM Mis-specified

	MSE						MAE					
	avg	.05	.25	.50	.75	.95	avg	.05	.25	.50	.75	.95
Stock												
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												
DFM	0.235	0.105	0.157	0.212	0.288	0.445	0.379	0.258	0.317	0.368	0.430	0.531
Chow–Lin	0.236	0.119	0.169	0.221	0.284	0.407	0.383	0.273	0.330	0.377	0.430	0.512
Spline	0.764	0.580	0.702	0.772	0.836	0.910	0.698	0.606	0.666	0.703	0.734	0.770
K-filter	0.761	0.570	0.691	0.760	0.822	0.909	0.697	0.603	0.663	0.698	0.729	0.775
K-smoother	0.762	0.569	0.692	0.761	0.823	0.910	0.698	0.601	0.663	0.698	0.730	0.774
Fraction of cases where balanced panel works better than non-balanced panel: 0.807												
Missing observations 40%												
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%												
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 (12) but with 10 factors in the DGP and 5 used in the factor model.

In the final set of experiments we focus on backdating 5% of missing observations when the variables are generated using the three DGPs in (12)–(14) but both y and the variables in X are subject to larger errors at the beginning of the sample. Specifically, an error term $5u$ is added to the first 5% of observations on y and each of the X variables, where u is i.i.d. standard normal. This in an interesting case since larger measurement errors at the beginning of the sample can be expected for some macroeconomic time series, consider for example backdating data for the unified Germany or the new European member states. Due to the stationarity of the variables, the results are also valid for the case of missing observations at the end of the sample, which is again frequent for macroeconomic time series where some variables are not timely released or are subject to subsequent revision. For these

Table 6
Disaggregation error, DGM is Chow–Lin, y depends on all Xs

	MSE						MAE					
	avg	.05	.25	.50	.75	.95	avg	.05	.25	.50	.75	.95
Stock												
DFM	0.004	0.002	0.003	0.004	0.005	0.007	0.043	0.031	0.037	0.042	0.047	0.058
Chow–Lin	0.055	0.029	0.039	0.051	0.065	0.095	0.160	0.117	0.138	0.156	0.178	0.214
Spline	1.024	0.648	0.858	1.011	1.181	1.445	0.692	0.552	0.636	0.691	0.750	0.825
K-filter	0.867	0.554	0.696	0.800	0.913	1.133	0.628	0.510	0.577	0.622	0.665	0.743
K-smoother	1.597	0.543	0.689	0.796	0.913	1.179	0.636	0.507	0.573	0.619	0.666	0.749
Fraction of cases where balanced panel works better than non-balanced panel: 0.535												
Flow												
DFM	0.004	0.002	0.003	0.004	0.004	0.006	0.048	0.035	0.042	0.048	0.054	0.064
Chow–Lin	0.053	0.026	0.038	0.048	0.062	0.094	0.180	0.129	0.155	0.176	0.200	0.247
Spline	0.537	0.349	0.452	0.532	0.618	0.727	0.584	0.469	0.539	0.586	0.632	0.692
K-filter	0.579	0.365	0.482	0.570	0.657	0.781	0.606	0.484	0.557	0.606	0.652	0.715
K-smoother	0.584	0.362	0.475	0.563	0.653	0.779	0.603	0.480	0.553	0.602	0.651	0.714
Fraction of cases where balanced panel works better than non-balanced panel: 0.429												
Missing observations 40%												
DFM	0.002	0.001	0.001	0.002	0.002	0.003	0.022	0.015	0.019	0.021	0.025	0.030
Chow–Lin	0.023	0.012	0.017	0.021	0.028	0.040	0.076	0.054	0.065	0.074	0.085	0.103
K-smoother	0.962	0.259	0.361	0.446	0.566	0.891	0.376	0.257	0.306	0.344	0.390	0.506
Fraction of cases where balanced panel works better than non-balanced panel: 0.339												
Missing observations 5%												
DFM	0.000	0.000	0.000	0.000	0.000	0.001	0.003	0.001	0.002	0.003	0.003	0.005
Chow–Lin	0.003	0.001	0.001	0.002	0.003	0.007	0.009	0.004	0.007	0.009	0.011	0.016
K-smoother	0.045	0.009	0.021	0.036	0.059	0.109	0.038	0.017	0.027	0.036	0.047	0.063
Fraction of cases where balanced panel works better than non-balanced panel: 0.500												

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 $Z_t = X_t$.

DGPs we would expect an improved performance of the univariate methods, since the use of regressors is complicated by the measurement error. Table 9 shows indeed that the range of the losses across interpolation methods shrinks substantially with respect to the standard case in Tables 2–4. However, even in this case, the multivariate methods perform quite well, either better than the univariate approaches or only slightly worse.

In summary, these additional experiments provide information on the situations where the factor method can be expected to perform best. The requirements include: a large number of explanatory variables for y, the variable to be interpolated or backdated; a limited idiosyncratic component for X, the set of variables to be used for factor extraction; and a limited measurement error for both y and X in those periods when y cannot be observed. On the other hand, if either y depends on few

Table 7
Disaggregation error, DGM is DFM, more variance idiosyncratic error

	MSE						MAE					
	avg	.05	.25	.50	.75	.95	avg	.05	.25	.50	.75	.95
Stock												
DFM	0.343	0.155	0.233	0.315	0.420	0.634	0.396	0.271	0.333	0.388	0.450	0.553
Chow–Lin	0.469	0.246	0.350	0.438	0.562	0.778	0.467	0.344	0.411	0.459	0.519	0.608
Spline	1.162	0.726	0.970	1.159	1.340	1.624	0.735	0.581	0.679	0.738	0.797	0.869
K-filter	0.872	0.608	0.747	0.844	0.962	1.182	0.643	0.534	0.598	0.638	0.682	0.757
K-smoother	0.919	0.591	0.742	0.845	0.970	1.233	0.646	0.530	0.595	0.638	0.684	0.772
Fraction of cases where balanced panel works better than non-balanced panel: 0.932												
Flow												
DFM	0.337	0.153	0.230	0.312	0.412	0.630	0.455	0.310	0.384	0.445	0.514	0.633
Chow–Lin	0.460	0.239	0.341	0.438	0.551	0.771	0.535	0.387	0.466	0.532	0.593	0.706
Spline	0.621	0.395	0.532	0.629	0.715	0.818	0.628	0.502	0.583	0.635	0.679	0.732
K-filter	0.644	0.434	0.559	0.647	0.728	0.833	0.640	0.525	0.597	0.643	0.684	0.735
K-smoother	0.642	0.430	0.556	0.645	0.725	0.834	0.638	0.522	0.596	0.642	0.683	0.736
Fraction of cases where balanced panel works better than non-balanced panel: 0.950												
Missing observations 40%												
DFM	0.170	0.073	0.114	0.154	0.208	0.322	0.205	0.137	0.170	0.199	0.233	0.293
Chow–Lin	0.217	0.111	0.158	0.204	0.262	0.369	0.233	0.167	0.201	0.229	0.260	0.311
K-smoother	1.728	0.278	0.367	0.450	0.547	0.990	0.392	0.265	0.307	0.342	0.380	0.533
Fraction of cases where balanced panel works better than non-balanced panel: 0.843												
Missing observations 5%												
DFM	0.021	0.004	0.010	0.017	0.027	0.053	0.026	0.012	0.018	0.024	0.031	0.044
Chow–Lin	0.026	0.005	0.014	0.022	0.035	0.062	0.029	0.013	0.021	0.028	0.035	0.048
K-smoother	0.048	0.010	0.024	0.041	0.063	0.113	0.039	0.019	0.028	0.038	0.048	0.065
Fraction of cases where balanced panel works better than non-balanced panel: 0.734												

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 (12), but the error term in the factor model is $2^{*}e_t$.

variables or the idiosyncratic component of X is substantial, Chow–Lin typically outperforms the factor approach. Finally, when there are substantial measurement errors, the univariate methods become reasonable competitors.

4. Empirical applications

In this section we compare the relative merits of the interpolation methods in practice using two datasets for, respectively, some European countries and the euro area.

In the first application, we consider quarterly series for GDP growth and inflation (measured as the quarter on quarter change in the private consumption deflator) for

Table 8
Disaggregation error, DGM is Chow–Lin, more variance idiosyncratic error

	MSE						MAE					
	avg	.05	.25	.50	.75	.95	avg	.05	.25	.50	.75	.95
Stock												
DFM	0.430	0.235	0.330	0.408	0.514	0.689	0.448	0.336	0.395	0.443	0.498	0.578
Chow–Lin	0.217	0.054	0.107	0.176	0.277	0.532	0.305	0.162	0.229	0.290	0.366	0.507
Spline	1.259	0.855	1.090	1.240	1.421	1.700	0.767	0.635	0.715	0.767	0.819	0.895
K-filter	1.229	0.665	0.784	0.868	0.969	1.168	0.659	0.557	0.610	0.646	0.683	0.753
K-smoother	1.281	0.662	0.792	0.878	0.986	1.245	0.666	0.558	0.613	0.648	0.689	0.776
Fraction of cases where balanced panel works better than non-balanced panel: 0.924												
Flow												
DFM	0.409	0.224	0.316	0.391	0.485	0.652	0.505	0.378	0.447	0.499	0.560	0.647
Chow–Lin	0.233	0.058	0.120	0.201	0.307	0.546	0.368	0.192	0.277	0.357	0.444	0.585
Spline	0.673	0.490	0.604	0.679	0.749	0.840	0.654	0.558	0.618	0.658	0.693	0.737
K-filter	0.692	0.500	0.618	0.695	0.761	0.859	0.663	0.563	0.626	0.668	0.701	0.746
K-smoother	0.690	0.500	0.616	0.694	0.761	0.861	0.662	0.563	0.625	0.667	0.701	0.747
Fraction of cases where balanced panel works better than non-balanced panel: 0.957												
Missing observations 40%												
DFM	0.211	0.113	0.157	0.200	0.256	0.347	0.230	0.168	0.201	0.226	0.257	0.302
Chow–Lin	0.076	0.021	0.038	0.060	0.095	0.183	0.133	0.073	0.099	0.125	0.158	0.219
K-smoother	1.732	0.288	0.376	0.438	0.521	1.172	0.403	0.272	0.309	0.338	0.370	0.574
Fraction of cases where balanced panel works better than non-balanced panel: 0.841												
Missing observations 5%												
DFM	0.026	0.005	0.013	0.022	0.034	0.061	0.028	0.013	0.021	0.027	0.035	0.048
Chow–Lin	0.009	0.001	0.003	0.006	0.011	0.026	0.016	0.006	0.010	0.015	0.020	0.031
K-smoother	0.051	0.011	0.026	0.043	0.067	0.114	0.040	0.019	0.030	0.039	0.049	0.066
Fraction of cases where balanced panel works better than non-balanced panel: 0.790												

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 the error term in the factor model is $2^* e_t$.

Austria, France, Finland, Germany, Italy, Spain and the Netherlands, over the period 1977:3–1999:2.[4] 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

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

Table 9
Backdating with measurement error, 5% missing obs

	MSE						MAE					
	avg	.05	.25	.50	.75	.95	avg	.05	.25	.50	.75	.95
DGM AR												
DFM	35.698	17.629	25.767	33.565	43.031	61.181	4.737	3.331	4.098	4.674	5.279	6.419
Chow–Lin	35.639	17.568	25.754	33.553	42.800	61.123	4.734	3.326	4.096	4.669	5.274	6.405
K-smoother	35.691	17.624	25.811	33.539	43.020	61.126	4.736	3.328	4.096	4.669	5.280	6.398
Fraction of cases where balanced panel works better than non-balanced panel: 0.765												
DGM DFM												
DFM	1.544	0.342	0.847	1.349	1.990	3.406	0.922	0.432	0.702	0.902	1.115	1.491
Chow–Lin	1.430	0.291	0.725	1.211	1.877	3.332	0.902	0.409	0.672	0.880	1.100	1.481
K-smoother	1.448	0.291	0.731	1.241	1.899	3.397	0.905	0.404	0.674	0.889	1.103	1.496
Fraction of cases where balanced panel works better than non-balanced panel: 0.959												
DGM Chow–Lin												
DFM	4.172	0.589	1.679	3.175	5.546	11.303	1.498	0.584	1.014	1.425	1.874	2.699
Chow–Lin	4.154	0.560	1.642	3.182	5.463	11.390	1.496	0.576	1.011	1.427	1.863	2.706
K-smoother	4.172	0.530	1.598	3.189	5.510	11.524	1.497	0.573	1.012	1.419	1.871	2.722
Fraction of cases where balanced panel works better than non-balanced panel: 0.946												

Note: The table reports the mean and percentiles of the empirical distribution of the MSE and MAE, computed over 2000 replications, when the first 5% of the observations on y and each of the X are equal to their values plus 5^*u, with u i.i.d. N(0, 1).

deflator, export and import deflators, etc., overall 50 series. For GDP growth, we use a set of real variables, that includes among 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–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 10.

As regards the interpolation of missing infra-year data, in the case of the inflation rates, the Chow–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 results are more varied: factor, Chow–Lin and the Kalman filter are best for two countries each, the Spline for one. When estimating missing

Table 10
Estimation of quarterly data

	MSE							MAE						
	AT	DE	ES	FI	FR	IT	AW	AT	DE	ES	FI	FR	IT	AW
Inflation														
DFM	0.42	0.47	0.13	0.31	0.09	0.06	0.19	0.43	0.46	0.26	0.36	0.19	0.16	0.30
Chow–Lin	0.47	0.25	0.11	0.23	0.06	0.02	0.20	0.44	0.32	0.20	0.33	0.15	0.10	0.30
Spline	0.55	0.71	0.28	0.34	0.18	0.07	0.24	0.49	0.55	0.33	0.39	0.27	0.18	0.31
K-filter	0.57	0.60	0.34	0.50	0.18	0.12	0.24	0.50	0.49	0.40	0.46	0.31	0.24	0.32
K-smoother	0.54	0.58	0.30	0.50	0.17	0.07	0.24	0.47	0.48	0.37	0.46	0.30	0.17	0.32
Missing observations 20%														
DFM	0.12	0.12	0.08	0.07	0.06	0.03	0.07	0.12	0.12	0.10	0.09	0.10	0.06	0.09
Chow–Lin	0.07	0.11	0.04	0.05	0.04	0.009	0.05	0.09	0.12	0.07	0.08	0.07	0.04	0.07
K-smoother	0.15	0.46	0.17	0.14	0.71	0.30	0.93	0.15	0.25	0.16	0.15	0.29	0.18	0.41

	MSE								MSE							
	AT	DE	ES	FI	FR	IT	NL	AW	AT	DE	ES	FI	FR	IT	NL	AW
Real GDP growth																
DFM	0.74	0.72	0.32	0.77	0.38	0.53	0.65	0.20	0.62	0.63	0.45	0.68	0.50	0.55	0.57	0.33
Chow–Lin	0.74	0.53	0.43	0.81	0.40	0.39	0.73	0.17	0.63	0.57	0.53	0.70	0.50	0.48	0.64	0.31
Spline	0.87	0.84	0.27	0.76	0.46	0.57	0.72	0.56	0.67	0.68	0.40	0.67	0.56	0.57	0.60	0.53
K-filter	0.76	0.80	0.26	0.83	0.48	0.63	0.58	0.47	0.62	0.69	0.40	0.72	0.59	0.62	0.55	0.51
K-smoother	0.78	0.86	0.28	0.79	0.51	0.63	0.58	0.46	0.63	0.71	0.41	0.69	0.62	0.62	0.55	0.51
Missing observations 20%																
DFM	0.40	0.14	0.29	0.35	0.23	0.28	0.28	0.10	0.21	0.14	0.21	0.20	0.20	0.20	0.18	0.12
Chow–Lin	0.38	0.12	0.22	0.24	0.22	0.19	1.21	0.01	0.20	0.14	0.19	0.18	0.19	0.16	0.40	0.05
K-smoother	0.46	0.44	0.34	0.26	0.31	0.43	0.35	0.23	0.23	0.27	0.22	0.20	0.21	0.26	0.20	0.19

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, AW: Area wide.

observations concentrated at the beginning of the sample, multivariate methods are better than univariate methods, and Chow–Lin is the best for the majority of the countries.

Overall, these results are in line with the outcome of the simulation experiments. In fact, we are grouping variables for rather different countries into a single dataset for factor extraction, so that we can expect the idiosyncratic component to be substantial, and we have seen that in this case Chow–Lin tends to outperform the factor method. Also, measurement error can be considered to be a more serious issue for GDP growth than for inflation, and indeed the univariate methods perform better for the former than for the latter.

To consider the issue of the homogeneity of the dataset, in the second application we repeat the interpolation procedure for inflation and GDP growth over the period 1980:1–2003:4, using 64 variables for the euro area extracted from an updated version of the dataset in Fagan et al. (2001), details are provided in the Data Appendix. Now the idiosyncratic component in the X variables can be expected to be smaller, and in fact the factor method works better than Chow–Lin for inflation and only slightly worse for GDP growth.

Finally, 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–Lin method; (b) decreased the number of regressors in the Chow–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.

5. Using the interpolated data

On 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 actual macroeconomic data employed in the previous section.

For the simulation experiments, we generate the data according to the three DGPs in (12)–(14). Then we compute the difference (ρ) between the first-order autocorrelation coefficients for the actual and interpolated series, and the absolute value of the difference (β) of the estimated coefficient of y_t in the regression $w_t = y_t + u_t$, with u_t i.i.d. N(0,1), using actual and interpolated data for both w_t and y_t.

The results are reported in Tables 11–13 for the three 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 (5% or 40%). As before, we report both the mean and percentiles of the empirical distribution of ρ and β over 2000 replications.

Four main comments can be made. First, the ranking of the disaggregation methods in terms of bias reflects that of Tables 2–4, 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 ρ and β 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 2–4 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, when the DGP is multivariate, univariate interpolation procedures lead to major biases in the

Table 11
Properties of interpolated data, DGM is DFM

	ρ						β					
	avg	.05	.25	.50	.75	.95	avg	.05	.25	.50	.75	.95
Stock												
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	0.008	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												
DFM	0.151	0.011	0.057	0.124	0.209	0.398	0.090	0.007	0.033	0.071	0.127	0.234
Chow–Lin	0.143	0.010	0.052	0.117	0.205	0.369	0.092	0.006	0.034	0.076	0.129	0.240
Spline	0.781	0.432	0.639	0.787	0.921	1.101	0.075	0.005	0.029	0.061	0.106	0.199
K-filter	0.585	0.219	0.442	0.590	0.736	0.913	0.083	0.006	0.030	0.064	0.112	0.222
K-smoother	0.605	0.228	0.465	0.612	0.759	0.941	0.411	0.005	0.030	0.063	0.112	0.221
Missing observations 40%												
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%												
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 (ρ) between the first-order autocorrelation coefficients for the actual and interpolated series, and the absolute value of the difference (β) of the estimated coefficient of y_t in the regression $w_t = y_t + u_t$, with u_t i.i.d. N(0, 1), using actual and interpolated data for both w_t and y_t. The DGP for y and X is as in (12).

estimated dynamics, i.e. large values of ρ, and viceversa when the DGP is univariate. Finally, even when ρ is large, the corresponding value of β is small, 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 actual macroeconomic data is concerned, we compute ρ as before, while β is the difference of the estimated coefficients in a regression of inflation or GDP growth for country i on the same variable for Germany, using actual and interpolated series. For the euro area data, β is computed as the difference of the estimated coefficient in a regression of inflation on GDP growth, and viceversa.

The results are summarized in Table 14 and three main comments are in order. First, for inflation the lowest values for ρ are achieved by the factor method in 5 out of 7 cases, with Chow–Lin being the best in the remaining two cases. On the other hand, Chow–Lin generates the lowest values for β in 2 out of 6 cases, with a univariate method performing best in the other four cases. The biases are in general

Table 12
Properties of interpolated data, DGM is Chow–Lin

	ρ						β					
	avg	.05	.25	.50	.75	.95	avg	.05	.25	.50	.75	.95
Stock												
DFM	0.096	0.007	0.036	0.078	0.133	0.266	0.117	0.010	0.047	0.096	0.165	0.297
Chow–Lin	0.086	0.006	0.029	0.065	0.119	0.231	0.117	0.008	0.044	0.097	0.163	0.295
Spline	0.663	0.340	0.529	0.668	0.795	0.981	0.107	0.008	0.043	0.089	0.150	0.271
K-filter	0.320	0.031	0.138	0.284	0.460	0.739	0.167	0.012	0.065	0.139	0.239	0.402
K-smoother	0.358	0.037	0.169	0.330	0.516	0.793	0.256	0.012	0.068	0.148	0.254	0.433
Flow												
DFM	0.150	0.010	0.053	0.114	0.208	0.401	0.091	0.006	0.033	0.074	0.129	0.240
Chow–Lin	0.119	0.009	0.042	0.094	0.166	0.322	0.096	0.006	0.038	0.078	0.134	0.247
Spline	0.778	0.484	0.657	0.779	0.894	1.073	0.077	0.006	0.029	0.063	0.110	0.196
K-filter	0.585	0.276	0.459	0.585	0.708	0.897	0.083	0.007	0.031	0.067	0.115	0.208
K-smoother	0.605	0.289	0.479	0.604	0.733	0.923	0.083	0.007	0.032	0.066	0.115	0.208
Missing observations 40%												
DFM	0.038	0.002	0.009	0.024	0.056	0.119	0.037	0.001	0.009	0.024	0.054	0.112
Chow–Lin	0.031	0.001	0.008	0.019	0.045	0.101	0.037	0.002	0.009	0.024	0.055	0.111
K-smoother	0.090	0.002	0.015	0.045	0.110	0.327	0.082	0.002	0.011	0.027	0.064	0.252
Missing observations 5%												
DFM	0.013	0.001	0.004	0.009	0.018	0.039	0.013	0.001	0.004	0.010	0.018	0.034
Chow–Lin	0.011	0.001	0.004	0.008	0.015	0.036	0.013	0.001	0.004	0.010	0.018	0.034
K-smoother	0.021	0.001	0.006	0.015	0.029	0.063	0.014	0.001	0.005	0.010	0.020	0.038

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

small for all interpolation procedures, ranging for ρ between 0.001 and 0.193, and for β between 0.001 and 0.133. Second, for GDP growth Chow–Lin is the best both in terms of ρ (6 out of 8 cases) and of β (5 out of 7 cases). The interesting result is that now the biases are larger, in the range 0.02–0.94 for ρ and 0.019–0.31 for β, with the larger biases in ρ associated with the univariate methods. This is presumably related to the lower persistence and higher infra-annual dynamics of GDP growth with respect to the inflation rate. Third, for the case of missing observations at the beginning of the sample, Chow–Lin is clearly the best as regards β for inflation, while the factor method performs better for GDP growth. The results are evenly distributed across methods for ρ. Both biases, for both variables, are in general smaller than in the case of interpolation.

In summary, for these time series the relative ranking of Chow–Lin and the factor method in terms of bias in subsequent econometric analyses with interpolated data is not clear cut. Yet, typically the multivariate interpolation methods induce lower biases in the estimated dynamics than the univariate procedures, and usually the bias in the parameters of contemporaneous relationships across variables are also

Table 13
Properties of interpolated data, DGM is AR(1)

	ρ						β					
	avg	.05	.25	.50	.75	.95	avg	.05	.25	.50	.75	.95
Stock												
DFM	0.602	0.237	0.486	0.635	0.741	0.868	0.134	0.010	0.054	0.113	0.190	0.334
Chow–Lin	0.461	0.144	0.316	0.450	0.597	0.793	0.186	0.017	0.086	0.167	0.270	0.426
Spline	0.093	0.010	0.047	0.085	0.128	0.204	0.107	0.008	0.041	0.088	0.153	0.271
K-filter	0.296	0.015	0.106	0.254	0.439	0.678	0.179	0.015	0.075	0.156	0.262	0.417
K-smoother	0.243	0.014	0.077	0.178	0.368	0.640	0.210	0.014	0.079	0.168	0.276	0.451
Flow												
DFM	0.191	0.012	0.055	0.135	0.283	0.539	0.059	0.005	0.021	0.047	0.083	0.154
Chow–Lin	0.383	0.054	0.227	0.377	0.527	0.734	0.118	0.009	0.041	0.092	0.170	0.315
Spline	0.161	0.073	0.122	0.157	0.196	0.262	0.036	0.003	0.014	0.030	0.051	0.090
K-filter	0.092	0.007	0.037	0.076	0.119	0.227	0.049	0.003	0.017	0.037	0.067	0.130
K-smoother	0.094	0.009	0.039	0.077	0.125	0.229	0.193	0.003	0.017	0.036	0.065	0.132
Missing observations 40%												
DFM	0.040	0.003	0.014	0.031	0.054	0.112	0.048	0.004	0.019	0.038	0.067	0.125
Chow–Lin	0.074	0.006	0.028	0.061	0.106	0.192	0.055	0.004	0.021	0.045	0.078	0.142
K-smoother	0.046	0.004	0.019	0.038	0.063	0.115	0.477	0.004	0.022	0.050	0.089	0.224
Missing observations 5%												
DFM	0.010	0.001	0.003	0.007	0.013	0.031	0.012	0.001	0.004	0.009	0.016	0.035
Chow–Lin	0.011	0.001	0.003	0.007	0.014	0.033	0.012	0.001	0.004	0.009	0.017	0.035
K-smoother	0.010	0.001	0.004	0.008	0.014	0.027	0.015	0.001	0.005	0.010	0.020	0.045

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

smaller. The good performance of the Chow–Lin procedure is likely due to the covariance structure of the datasets, whereby some variables are presumably highly correlated both at the disaggregate and at the aggregate level with the series to be interpolated, as in the case of the DGP (13). In this context, the variable selection procedure implemented for the Chow–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. Finally, 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.

6. 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

Table 14
Properties of interpolated data, empirical example

	ρ							β					
	AT	DE	ES	FI	FR	IT	AW	AT	ES	FI	FR	IT	AW
Inflation													
DFM	0.161	0.086	0.003	0.065	0.027	0.001	0.003	0.089	0.047	0.021	0.019	0.002	0.085
Chow–Lin	0.116	0.101	0.004	0.068	0.023	0.006	0.017	0.035	0.016	0.016	0.025	0.001	0.013
Spline	0.183	0.183	0.030	0.083	0.037	0.009	0.048	0.101	0.049	0.010	0.015	0.017	0.005
K-filter	0.174	0.169	0.031	0.078	0.026	0.005	0.042	0.133	0.011	0.037	0.017	0.010	0.076
K-smoother	0.185	0.193	0.032	0.079	0.026	0.011	0.042	0.129	0.020	0.040	0.013	0.013	0.077
Missing observations 20%													
DFM	0.040	0.061	0.028	0.031	0.012	0.002	0.027	0.020	0.056	0.053	0.079	0.020	0.411
Chow–Lin	0.004	0.029	0.015	0.026	0.011	0.004	0.008	0.008	0.004	0.007	0.017	0.014	0.071
K-smoother	0.002	0.018	0.026	0.028	0.000	0.008	0.045	0.160	0.117	0.134	0.237	0.128	0.411

	ρ								β						
	AT	DE	ES	FI	FR	IT	NL	AW	AT	ES	FI	FR	IT	NL	AW
Real GDP growth															
DFM	0.18	0.53	0.32	0.21	0.38	0.45	0.20	0.297	0.09	0.24	0.27	0.09	0.008	0.34	0.204
Chow–Lin	0.54	0.33	0.06	0.60	0.02	0.20	0.12	0.025	0.26	0.06	0.05	0.05	0.12	0.24	0.019
Spline	0.84	0.94	0.08	0.88	0.31	0.42	0.64	0.346	0.17	0.19	0.24	0.13	0.17	0.30	0.211
K-filter	0.77	0.84	0.07	0.84	0.26	0.31	0.57	0.296	0.18	0.19	0.23	0.009	0.18	0.31	0.235
K-smoother	0.77	0.85	0.08	0.87	0.29	0.32	0.57	0.297	0.18	0.19	0.24	0.03	0.18	0.31	0.235
Missing observations 20%															
DFM	0.23	0.08	0.04	0.05	0.06	0.05	0.19	0.097	0.18	0.04	0.002	0.05	0.25	0.05	0.005
Chow–Lin	0.26	0.05	0.04	0.08	0.03	0.05	0.03	0.016	0.19	0.09	0.031	0.05	0.23	0.04	0.014
K-smoother	0.20	0.02	0.04	0.07	0.07	0.05	0.20	0.082	0.26	0.06	0.012	0.07	0.25	0.06	0.044

Note: The table reports the difference (ρ) between the first-order autocorrelation coefficients for the actual and interpolated series, and the absolute value of the difference (β) of the estimated coefficients in a regression of inflation or GDP growth for country i on the same variable for Germany, using actual and interpolated series.
AT: Austria, DE: Germany, ES: Spain, FI: Finland, FR: France, IT: Italy, NL: The Netherlands, AW: Area wide.

large datasets, 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, although 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–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) models for each variable with correlated errors, or factor models where the variable of interest can depend either on the factors or on a small subset of other variables. As a sensitivity analysis, we have considered factor models comprising a number of factors largely inferior to that of the DGP, or with larger idiosyncratic components, or with variables subject to measurement error. 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 the factor method can be expected to perform best when there is a large number of explanatory variables for y, the variable to be interpolated or backdated; a limited idiosyncratic component for X, the set of variables to be used for factor extraction; and a limited measurement error for both y and X in those periods when y cannot be observed. If either y depends on few variables or the idiosyncratic component of X is substantial, Chow–Lin typically outperforms the factor approach. Finally, when there are substantial measurement errors, the univariate methods become reasonable competitors.

Third, we have used actual time-series on quarterly GDP growth rate and inflation for 7 European countries and for 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 main reason underlying the good results for the Chow and Lin technique is the pre-selection of the regressors according to their 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, 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 again favor the multivariate approaches, with a less clear-cut ranking of Chow–Lin and the factor method in the empirical analysis. An interesting caveat resulting from this part of the paper is that biases can be sizeable, especially in the case of interpolation where there are a relatively large number of missing observations. Non-standard limit distributions might also be required for inference when using interpolated data, and this provides an interesting topic for further research in this area.

7. Uncited reference

Chan (1993).

Acknowledgements

We are grateful to four anonymous Referees, Günter Coenen, Lutz Kilian, Jim Stock and participants in the CEPR Euro Area Business Cycle Network workshop at Bocconi University, the 2002 CEFI conference in Aix en Provence and an ECB seminar for helpful comments on a previous version. The usual disclaimers apply.

Appendix. Data

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	

List of variables in area wide dataset

CAN	current account balance
COMPR	commodity prices
EEN	effective exchange rate
EER	effective exchange rate
GCD	government consumption deflator
GCR	government consumption
GDN_YEN	ratio public debt/GDP
GIN	public investment
GIN_OTHER	public investment other
GIX	implicit public debt interest rate
GLN_YEN	ratio government net lending/GDP
GON	gross operating surplus
GPN_YEN	ratio government primary surplus/GDP
GRN_YEN	ratio government revenue/GDP
GSN_YEN	government savings/GDP
GYN_YEN	government disposable income/GDP
HICP	HICP (NSA)
INN_YEN	government interest payments/GDP
ITD	gross investment deflator
ITR	gross investment
KSR	whole-economy capital stock
LEN	employees (persons)
LFN	labour force
LNN	total employment (persons)
LPROD	labour productivity
LTN	long-term interest rate

MTD	imports of goods and services deflator
MTR	imports of goods and services
NFA	net foreign assets
NFN	net factor income from abroad
PCD	consumption deflator
PCR	consumption
PYR	household's disposable income
SCD	variation of stocks deflator
SCR	variation of stocks
SSN_YEN	social security contributions total/GDP
STN	short-term interest rate (nominal in percent)
STRQ	short-term quarterly interest rate (Real)
TBR	trade balance
TDN_YEN	direct taxes excluding social security contributions/GDP
TIN_YEN	ratio, indirect taxes/GDP
TKN_YEN	total revenue minus current revenue/GDP
TRN_FIRMS_YEN	transfers to firms/GDP
TRN_OTHER_YEN	other transfers/GDP
TRN_YEN	transfers/GDP
ULC	unit labour costs
ULT	tend unit labour costs
UNN	number of unemployed
URT	trend unemployment
URX	unemployment
WIN	compensation to employees
WLR	wealth
WRN	wage rate
XTD	exports of goods and services deflator
XTR	exports of goods and services
YED	GDP deflator
YER	GDP
YET	potential output
YFD	GDP at factor costs deflator
YGA	Output gap
YWD	World GDP deflator
YWDX	World demand deflator, composite indicator
YWR	World GDP
YWRX	World demand, composite indicator

References

Angelini, E., Henry, J., Mestre, R., 2001. Diffusion index based inflation forecasts for the Euro area. European Central Bank WP 61.

Angelini, E., Henry, J., Marcellino, M., 2003. Interpolation with a large information set. ECB WP 252.

Artis, M., Banerjee, A., Marcellino, M., 2004. Factor forecasts for the UK. Journal of Forecasting, forthcoming.

Bai, J., 2003. Inferential theory for factor models of large dimension. Econometrica 71, 135–171.

Bai, J., Ng, S., 2004. Confidence intervals for diffusion index forecasts with a large number of predictors. Mimeo.

Banerjee, A., Marcellino, M., Masten, I., 2005. Forecasting macroeconomic variables for the accession countries. In: Artis, M., Banerjee, A., Marcellino, M. (Eds.), The European Enlargement: Prospects and Challenges. Cambridge University Press, Cambridge, forthcoming.

Boivin, J., Ng, S., 2004. Are more data always better for factor analysis? Journal of Econometrics, forthcoming.

Boot, J.C.G., Feibes, W., Lisman, J.H.C., 1967. Further methods of derivation of quarterly figures from annual data. Applied Statistics 16, 65–75.

Chamberlain, G., Rothschild, M., 1983. Arbitrage factor structure, and mean variance analysis of large asset markets. Econometrica 51, 1281–1304.

Chan, W.S., 1993. Disaggregation of annual time-series data into quarterly figures: a comparative study. Journal of Forecasting 12, 677–688.

Chow, G.C., Lin, A., 1971. Best linear unbiased interpolation distribution and extrapolation of time-series by related series. Review of Economics and Statistics 53, 372–375.

Cohen, K.J., Müller, W., Padberg, M.W., 1971. Autoregressive approaches to disaggregation of time series data. Applied Statistics 20, 119–129.

Connor, G., Korajczyk, R.A., 1986. Performance measurement with the arbitrage pricing theory. Journal of Financial Economics 15, 373–394.

Connor, G., Korajczyk, R.A., 1993. A test for the number of factors in an approximate factor model. Journal of Finance 48, 1263–1291.

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.

Fagan, G., Henry, J., Mestre, R., 2001. An area-wide model for the Euro area. ECB Working Paper No. 42, European Central Bank.

Fernandez, R.B., 1981. A methodological note on the estimation of time-series. Review of Economics and Statistics 63, 471–476.

Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2000. The generalized factor model: identification and estimation. The Review of Economic and Statistics 82, 540–554.

Gomez, V., Maravall, A., 1994. Estimation, prediction and interpolation for non stationary time series with the Kalman filter. Journal of the American Statistical Association 89, 611–624.

Guerrero, V.M., 1990. Temporal disaggregation of time-series: an ARIMA-based approach. International Statistical Review 58, 29–46.

Harvey, A.C., Pierse, R.G., 1984. Estimating missing observations in economic time series. Journal of the American Statistical Association 79, 125–131.

Kohn, R., Ansley, C.F., 1986. Estimation, prediction and interpolation for ARIMA models with missing data. Journal of the American Statistical Association 81, 751–761.

Lisman, J.H.C., Sandee, J., 1964. Derivation of quarterly figures from annual data. Applied Statistics 13, 87–90.

Litterman, R.B., 1983. A random walk, Markov model for the distribution of time series. Journal of Business and Economic Statistics 1, 169–173.

Marcellino, M., 1998. Temporal disaggregation, missing observations, outliers, and forecasting: a unifying non-model based procedure. Advances in Econometrics 13, 181–202.

Marcellino, M., Stock, J.H., Watson, M.W., 2003. Macroeconomic forecasting in the Euro area: country specific versus euro wide information. European Economic Review 47, 1–18.

Micula, G., Micula, S., 1998. Handbook of Splines. Kluwer Academic Publishers, Dordrecht.

Nijman, T.E., Palm, F.C., 1986. The construction and use of approximations for missing quarterly observations: a model-based approach. Journal of Business and Economic Statistics 4, 47–58.

Stock, J.H., Watson, M.W., 2002a. Macroeconomic forecasting using diffusion indexes. Journal of Business and Economic Statistics 20, 147–162.

Stock, J.H., Watson, M.W., 2002b. Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association 97, 1167–1179.

Stram, D.O., Wei, W.W.S., 1986. A methodological note on disaggregation of time series totals. Journal of Time Series Analysis 7, 293–302.

Wei, W.W.S., Stram, D.O., 1990. Disaggregation of time series models. Journal of the Royal Statistical Society, Series B 52, 453–467.