

Automatic Methods for the Back Recalculation of Economic Data

Andrew P. Blake

National Institute of Economic and Social Research
2 Dean Trench Street, Smith Square, London SW1P 3HE
ablake@niesr.ac.uk

and

Richard G. Pierse

University of Surrey and NIESR
r.pierse@surrey.ac.uk

December 16, 2002

Abstract

In this paper we discuss methods for the retriangulation or automatic recalculation of data. This is typically needed when there are substantial revisions to data definitions. We consider the implication of using cointegration as an aid as well simpler linear interpolation methods. We conclude that simpler is better, particularly as it is often that fully revised data and unrevised data fail to cointegrate.

1 Introduction

In this paper we discuss the revision of national accounts data. It describes efficient available methods for the back-calculation or *retriangulation* of national accounts data in the face of an occasional large data revision. This is needed because of the introduction of the Euro at the end of 1998 as well as the introduction of ESA 1995 and plans for consistent accounts across all Eurozone countries. Some statistical offices have already begun such revision processes. For example, historical data for the Netherlands consistent with SNA standards was published in den Bakker, Huitker, and van Bochove (1990). In den Bakker and van Rooijen (1997, Appendix I) an account of seven different revision exercises undertaken by Statistics Netherlands is given, each of which used a different approach. A major conclusion of their paper is that the manpower required for a full revision was excessive, and simpler alternatives need to be

found, subject, of course, to their adequate performance. This paper focuses on suggested possible alternatives to a full revision of the national accounts data. In particular, it discusses largely automatic methods which can be straightforwardly applied to large quantities of quite disparate data on a consistent basis. Such methods must be flexible enough to take account of available additional information and handle both annual and quarterly accounts. The final object of this research is to determine appropriate techniques for this purpose, taking into account those already applied by statistical offices, culminating in the estimation of national accounts data.

There are two broad categories of approach for approximating a full revision of the national accounts. The first is to calculate from a reference point on a period-by-period basis, usually backwards in time. This can be done at different levels of detail, but at its simplest, the growth rates of the most closely related series are used to project the series backwards from the fully revised data. The calculation of a set of quarterly accounts as growth rates from such a reference point is detailed in van Hao and Buiten (1999) where they calculate accounts consistent with ESA 1995. This reduces the burden of a full revision by assuming that there are good enough indicator variables available from the existing definitions. However, choosing a different reference point could yield markedly different results.

The second approach is to interpolate revisions between a number of chosen reference years where in those years a full revision of the data is undertaken (see INSEE 1997, den Bakker and van Rooijen 1997). This could be seen as smoothing the growth rate approach so that the growth rates are consistent across revision years. We intend to focus on methods based on this general approach—interpolation based on reference years—as this offers the greater scope for the introduction of flexible automatic methods and uses more of the available information if more than one reference year is revised.

Similar difficulties exist when approaching the problem of temporal disaggregation and possible solutions are similar (see Chow and Lin 1971, Chow and Lin 1976, Harvey and Pierse 1984). We describe linear interpolation methods and Kalman filter methods and outline possible generalisations of both. We describe approaches which are largely ‘piecemeal’ to the calculation of each series in the sense that they then have to be subsequently reconciled to ensure the accounting constraints hold. Imposing those constraints as part of the

retrapolation procedure is an important generalisation of these procedures.

The choices for implementation of such a system are threefold. First, the chosen description of the unknown series with respect to the unrevised accounts. Second, the method of estimation which we refer to (loosely) as interpolation given the nature of the available data. Third, the imposition of both the balancing problem and the treatment of the quarterly and annual restrictions. We deal with each in turn.

For the first of these there are four possible approaches; the applicability of each depends on the available information. These correspond to the method of modelling the infrequently-observed time series. It can be modelled by its levels, its growth rate, its share of a known aggregate or its discrepancy from some other known series. For the second, the choice of estimation approach depends primarily on whether a parametric behavioural model is used to obtain the series. If this is the case, the model is naturally cast as containing unobserved variables in state-space and estimated by maximum likelihood methods. If not, then the discrepancy needs to be modelled to fit the reference years and interpolation methods applied.

If the method of estimating the revised data does not impose accounting constraints at the time of calculation, then they will need to be adjusted subsequently. The balancing problem requires some assessment of data reliabilities (see Sefton and Weale 1995). This will clearly depend on the original series as much as the new sets of accounts: reliabilities for the existing series should be readily obtainable from statistical offices, and it may be possible for them to be associated with the new series. If this is not the case, then an estimate of the data reliabilities will need to be made at the time of calculation. We do not deal with the balancing problem, although we discuss it further below.

In all that follows we assume that there exists some (possibly nonlinear and/or time-varying) relationship:

$$\mathbf{y}_t = f(\mathbf{x}_t) \approx \mathbf{A}_t \mathbf{x}_t \tag{1}$$

between a vector of observed variables at time t denoted by \mathbf{x}_t , and infrequently observed, redefined variables, \mathbf{y}_t . We write the model including the approximation above as any nonlinear model can be locally approximated using a time-varying linear model. We will usually discuss the problem with reference to the univariate case $y_t = \mathbf{a}'_t \mathbf{x}_t$, a single, representative row of (1).

Aggregate Eurozone data, rather than individual national account data, has also been analysed by Beyer, Doornik, and Hendry (2000, 2001). They describe a number of methods to improve the accuracy of revisions, including suggesting that they are better applied to the logarithms of some series and how best to deal with price indices. Much of their analysis is dependent on aggregation issues, which is beyond the scope of this discussion. We also leave aside the appropriate choice of benchmark years for the full revisions. These will clearly be years for which the most information is readily available, such as census years and historical rebasing years. It may be the outcome of any study of a retrapolation exercise to indicate a *minimum* number of reference years required for reliable estimates to be produced.

Finally, note that Liu and Hall (2001) consider a related problem, that of creating higher-frequency accounts than the data allows. They find that the simpler methods work almost as well as the more sophisticated ones. In our work we have found probably a more extreme result, where the simpler methods can often outperform more sophisticated ones.

The paper is organised as follows. We critically evaluate two general approaches in section 2. We begin in section 2.1 by discussing an approach which treats the retrapolated series as unobserved, latent variables. This relies on using the Kalman-Bucy filter to provide optimal estimates. The discussion illustrates the various assumptions that need to be made for this approach to be appropriate and indicates potential problems. In section 2.2 we then describe a conceptually simpler approach, based on linear interpolation, either in levels or as share equations. It emerges that there are great similarities between the two and we provide a synthesis.

In section 3 we evaluate retrapolation models by Monte Carlo simulation for appropriately chosen parameters. In section 3.1 we report the results of the Monte Carlo exercise on stationary data using Chow-Lin based interpolation. In section 3.2 we investigate cointegrated data using Kalman filter based methods. In section 3.3 we evaluate the results. In section 4 we discuss some properties of available data. We find that little of the data is cointegrated across definitions, which causes severe problems for the appropriate choice of model. Bearing in mind the results of all the previous sections, in section 5 we present some results for linear extrapolation of the data discussed in the previous section. Finally, in section 6 we draw conclusions.

2 Models considered

2.1 A Kalman Filter approach

The Kalman filter (Kalman 1960, Kalman and Bucy 1961) is an algorithm for providing optimal forecasts in any model that can be cast in state space form. The state space form is defined by two equations. The transition equation:

$$\mathbf{a}_t = \mathbf{T}_t \mathbf{a}_{t-1} + \mathbf{R}_t \mathbf{u}_t \quad (2)$$

relates the $n \times 1$ vector of state variables \mathbf{a}_t to their own past and to an error term \mathbf{u}_t . The state variables are not directly observed but are related to a vector of observed variables \mathbf{y}_t by a measurement equation:

$$\mathbf{y}_t = \mathbf{H}_t \mathbf{a}_t + \mathbf{v}_t. \quad (3)$$

The Kalman filter consists of a set of prediction equations to recursively calculate forecasts $\hat{\mathbf{a}}_{t|t-1}$ and a set of updating equations to update the prediction to $\hat{\mathbf{a}}_{t|t}$ whenever there is a new observation \mathbf{y}_t . When observations are missing, the updating equations are simply skipped and so the Kalman filter deals naturally with (irregularly occurring) missing observations. This would seem to make it a good choice as a procedure for reinterpolation. The parameters in the state matrices \mathbf{T}_t , \mathbf{R}_t and \mathbf{H}_t can be estimated by maximum likelihood using the prediction error decomposition of the likelihood function. Finally, smoothed estimators of the state variables, $\hat{\mathbf{a}}_{t|T}$, can be calculated, making use of all information in the sample $t = 1, \dots, T$. A clear exposition of the Kalman filter is provided in Harvey (1993, Chapter 4).

The use of the Kalman filter for interpolation problems based on economic data was described in Harvey and Pierse (1984), including techniques appropriate for non-stationary series. See also Gomez and Maravall (1994). In what follows we assume that appropriate likelihood maximisation is carried out on the basis of the models outlined. A brief representative overview is given in Harvey (1993, Chapter 3).

2.1.1 A model without cointegration

INSEE (1997) proposes two alternative state space models to be estimated within a Kalman filter framework. The most general version of the first model proposed by INSEE (1997), Model 1, is defined by the equations:

$$\Delta y_t = \alpha \Delta y_{t-1} + \beta' \Delta \mathbf{x}_{t-1} + \rho' \Delta \mathbf{x}_t + \varepsilon_t \quad (4)$$

and:

$$\Delta \mathbf{x}_t = \gamma \Delta y_{t-1} + \Phi \Delta \mathbf{x}_{t-1} + \eta_t \quad (5)$$

where \mathbf{x}_t is an $n \times 1$ vector of variables and y_t is a scalar.

The variable y_t is a variable in the new accounting system that we wish to retruncate over the past. \mathbf{x}_t is a vector of variables on the old accounting system that are observed over the past and for some period overlapping with the observations for y_t . The model is set out in differences. Reasons for this might be, either that it is assumed that the variables y_t and \mathbf{x}_t are both $I(1)$ so that first differencing is done to transform to stationarity, or simply that it is better to work with changes (growth rates if the variables are in logarithms) than with levels. Note that the first equation relates y_t to current (as well as lagged) values of \mathbf{x}_t . While it may be reasonable to assume that current values of \mathbf{x}_t may be more highly correlated with y_t than lagged values, including current values means that the model needs to be transformed to put it into a state space form before the Kalman filter can be used.

We assume that $\varepsilon_t \sim N(0, \sigma_1^2)$ and $\eta_t \sim N(0, \Sigma)$ where:

$$\Sigma = \begin{bmatrix} \sigma_2^2 & 0 & \cdots & 0 \\ 0 & \sigma_3^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p^2 \end{bmatrix}.$$

The diagonality of Σ imposes some restrictions on the quasi-VAR representation of $\Delta \mathbf{x}_t$. These restrictions are sufficient to identify the model parameters uniquely.

Note that the model can be rewritten as:

$$\begin{bmatrix} 1 & -\rho' \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \Delta \tilde{y}_t \\ \Delta \tilde{\mathbf{x}}_t \end{bmatrix} = \begin{bmatrix} \alpha & \beta' \\ \gamma & \Phi \end{bmatrix} \begin{bmatrix} \Delta \tilde{y}_{t-1} \\ \Delta \tilde{\mathbf{x}}_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ \eta_t \end{bmatrix}. \quad (6)$$

where tildas have been introduced to distinguish the latent state variables, which may or not be observed, from the actual data observations on them.

This can be written as the reduced form:

$$\begin{bmatrix} \Delta \tilde{y}_t \\ \Delta \tilde{\mathbf{x}}_t \end{bmatrix} = \begin{bmatrix} \alpha + \rho' \gamma & \beta' + \rho' \Phi \\ \gamma & \Phi \end{bmatrix} \begin{bmatrix} \Delta \tilde{y}_{t-1} \\ \Delta \tilde{\mathbf{x}}_{t-1} \end{bmatrix} + \begin{bmatrix} 1 & \rho' \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \varepsilon_t \\ \eta_t \end{bmatrix} \quad (7)$$

or:

$$\mathbf{z}_t = \mathbf{T} \mathbf{z}_{t-1} + \mathbf{R} \mathbf{e}_t \quad (8)$$

where $\mathbf{z}_t = (\Delta\tilde{y}_t, \Delta\tilde{\mathbf{x}}_t)'$ and $\mathbf{e}_t = (\varepsilon_t, \eta_t)'$. This is now in standard state space form. Note that the transformation introduces a covariance between the error terms ε_t and η_t . The parameters in this transformed equation are identified in the sense that they can be recovered uniquely from the state space form. The wider (and important) issue of identification in Kalman filter models is considered below.

The transition equation (7) needs to be augmented by a measurement equation, linking the state variables $\Delta\tilde{y}_t$ and $\Delta\tilde{\mathbf{x}}_t$, to actual observations Δy_t and $\Delta \mathbf{x}_t$. We assume that for the k observations, $t = t_{R_1}, \dots, t_{R_k}$, both Δy_t and $\Delta \mathbf{x}_t$ are observed and the subscript R indicates a revision period. These correspond to the possibly irregularly spaced revision dates. For these k periods the measurement equation is simply given by:

$$\begin{bmatrix} \Delta y_t \\ \Delta \mathbf{x}_t \end{bmatrix} = \begin{bmatrix} \Delta\tilde{y}_t \\ \Delta\tilde{\mathbf{x}}_t \end{bmatrix} = \mathbf{z}_t. \quad (9)$$

For the remaining $T - k$ observations (for example $t = 1, \dots, t_{R_1} - 1, t_{R_1} + 1, \dots, T - 1$ where, for purposes of exposition, we assume a revised set of accounts in the last period and t_{R_1} is not adjacent to any other revision date) only $\Delta \mathbf{x}_t$ is observed and the measurement equation is given by:

$$\Delta \mathbf{x}_t = \Delta\tilde{\mathbf{x}}_t. \quad (10)$$

For these latter observations, the Kalman filter updating equations for Δy_t are skipped and the best predictor of Δy_t is $\Delta\tilde{y}_t$ which is defined by:

$$\Delta\tilde{y}_t = \Delta\tilde{y}_{t|k} = [1 \quad \mathbf{0}_n] \mathbf{T} \begin{bmatrix} \Delta\tilde{y}_{t-1} \\ \Delta\tilde{\mathbf{x}}_{t-1} \end{bmatrix}. \quad (11)$$

Note that these measurement equations assume no measurement error so that the latent variables are simply equal to the observable variables, whenever there is an observation. Clearly, one could consider relaxing this restriction and allowing a measurement error as in the standard Kalman filter framework. We comment on this further below.

This model, Model 1, essentially relies on reinterpolation by the use of growth rates, in much the same way as a growth rate adjustment to a single reference year. The difference is that the known growth rates in the multiple reference years are satisfied exactly by the method. It does, of course, require that there are enough adjacent periods calculated for the reference periods to enable

growth rates to be used. With annual data and no adjacent periods, for example, this model would be inappropriate, such as is implicit in the dating scheme outlined above.

2.1.2 A cointegrated model

The second model considered by INSEE (1997), Model 2, is based on the assumption of a cointegrating vector linking the variables y_t and \mathbf{x}_t . This model could be appropriate if the underlying series are integrated. The model can be written as the pair of dynamic equations:

$$e_t = \alpha e_{t-1} + \beta' \Delta \mathbf{x}_{t-1} + \rho' \Delta \mathbf{x}_t + \varepsilon_t \quad (12)$$

and:

$$\Delta \mathbf{x}_t = \gamma e_{t-1} + \Phi \Delta \mathbf{x}_{t-1} + \eta_t \quad (13)$$

where $e_t = y_t - \nu' \mathbf{x}_t$ is the cointegrating vector which is assumed to be related to its own past and current and lagged values of \mathbf{x}_t . As with Model 1, the model needs to be transformed to put it into state-space form. Rewriting the equations we have:

$$\begin{bmatrix} 1 & -\rho' & \rho' \\ \mathbf{0} & \mathbf{I} & -\mathbf{I} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \tilde{e}_t \\ \tilde{\mathbf{x}}_t \\ \tilde{\mathbf{x}}_{t-1} \end{bmatrix} = \begin{bmatrix} \alpha & \beta' & -\beta' \\ \gamma & \Phi & -\Phi \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{e}_{t-1} \\ \tilde{\mathbf{x}}_{t-1} \\ \tilde{\mathbf{x}}_{t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ \eta_t \\ \mathbf{0} \end{bmatrix}$$

and as the reduced, state space form:

$$\begin{bmatrix} \tilde{e}_t \\ \tilde{\mathbf{x}}_t \\ \tilde{\mathbf{x}}_{t-1} \end{bmatrix} = \begin{bmatrix} \alpha + \rho' \gamma & \beta' + \rho' \Phi & -\beta' - \rho' \Phi \\ \gamma & \mathbf{I} + \Phi & -\Phi \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{e}_{t-1} \\ \tilde{\mathbf{x}}_{t-1} \\ \tilde{\mathbf{x}}_{t-2} \end{bmatrix} + \begin{bmatrix} 1 & \rho' \\ \mathbf{0} & \mathbf{I} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \varepsilon_t \\ \eta_t \end{bmatrix}. \quad (14)$$

The transition equation (14) is augmented by the measurement equation:

$$\begin{bmatrix} y_t \\ \mathbf{x}_t \end{bmatrix} = \begin{bmatrix} 1 & \nu' & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{e}_t \\ \tilde{\mathbf{x}}_t \\ \tilde{\mathbf{x}}_{t-1} \end{bmatrix} \quad (15)$$

where observations for both are available and:

$$\mathbf{x}_t = \tilde{\mathbf{x}}_t \quad (16)$$

where only \mathbf{x}_t is observed. Note that the parameters of the cointegrating vector, ν , appear only in the first measurement equation and so estimation of these parameters depends on sufficient overlap periods where both y_t and \mathbf{x}_t are observed.

This is quite an unusual model. What would normally be considered the cointegrating residual, in this case the discrepancy between the old and the new account, is modelled as an autoregressive process. Whilst it should be the case that the new and the old series should be cointegrated where appropriate, this is a non-standard way of modelling the residual. It is clearly not the usual error-correction form although the treatment of the corrections as the latent variable is quite natural.

2.1.3 An alternative model with cointegration

As we have just noted, the cointegrating model Model 2 is in some respects rather unusual. In particular it is not standard to model the cointegrating vector as a function of its own past. An alternative, more conventional, model can be constructed on the basis of an assumed cointegrating relationship between the $I(1)$ variables y_t and \mathbf{x}_t . From the Granger representation theorem, we are able to determine an appropriate modelling framework. This can be expressed through an error correction model (ECM) of the form:

$$\Delta y_t = \alpha \Delta y_{t-1} + \beta' \Delta \mathbf{x}_{t-1} + \tau (y_{t-1} - \nu' \mathbf{x}_{t-1}) + u_t. \quad (17)$$

The model is completed by assuming an unrestricted VAR process for the observable variables \mathbf{x}_t of the form:

$$\Delta \mathbf{x}_t = \Phi \Delta \mathbf{x}_{t-1} + \mathbf{v}_t. \quad (18)$$

This model can be set up in state space form with transition equation:

$$\begin{bmatrix} \tilde{y}_t \\ \tilde{\mathbf{x}}_t \\ \tilde{y}_{t-1} \\ \tilde{\mathbf{x}}_{t-1} \end{bmatrix} = \begin{bmatrix} 1 + \alpha + \tau & \beta' - \tau \nu' & -\alpha & -\beta' \\ \mathbf{0} & \mathbf{I} - \Phi & \mathbf{0} & -\Phi \\ 1 & \mathbf{0}' & 0 & \mathbf{0}' \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{y}_{t-1} \\ \tilde{\mathbf{x}}_{t-1} \\ \tilde{y}_{t-2} \\ \tilde{\mathbf{x}}_{t-2} \end{bmatrix} + \begin{bmatrix} u_t \\ \mathbf{v}_t \\ 0 \\ \mathbf{0} \end{bmatrix} \quad (19)$$

together with the measurement equation:

$$\begin{bmatrix} y_t \\ \mathbf{x}_t \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0}' & 0 & \mathbf{0}' \\ 0 & \mathbf{I} & 0 & \mathbf{0}' \end{bmatrix} \begin{bmatrix} \tilde{y}_t \\ \tilde{\mathbf{x}}_t \\ \tilde{y}_{t-1} \\ \tilde{\mathbf{x}}_{t-1} \end{bmatrix} \quad (20)$$

for the k periods of overlap and a similar appropriate observation equation for the $T - k$ remaining observations.

Note that if we generalize this to include contemporaneous \mathbf{x}_t in (17) we obtain:

$$\Delta y_t = \alpha \Delta y_{t-1} + \beta' \Delta \mathbf{x}_{t-1} + \rho' \Delta \mathbf{x}_t + \tau(y_{t-1} - \nu' \mathbf{x}_{t-1}) + u_t. \quad (21)$$

or, in state space:

$$\begin{bmatrix} \tilde{y}_t \\ \tilde{\mathbf{x}}_t \\ \tilde{y}_{t-1} \\ \tilde{\mathbf{x}}_{t-1} \end{bmatrix} = \begin{bmatrix} 1 + \alpha + \tau & \beta' + \rho' \Phi - \tau \nu' & -\alpha & -\beta' - \rho' \Phi \\ \mathbf{0} & \mathbf{I} - \Phi & \mathbf{0} & -\Phi \\ 1 & \mathbf{0}' & 0 & \mathbf{0}' \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{y}_{t-1} \\ \tilde{\mathbf{x}}_{t-1} \\ \tilde{y}_{t-2} \\ \tilde{\mathbf{x}}_{t-2} \end{bmatrix} + \begin{bmatrix} 1 & \rho' \\ \mathbf{0} & \mathbf{I} \\ 0 & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} u_t \\ \mathbf{v}_t \end{bmatrix} \quad (22)$$

with identical measurement equation (20) and so on.

This model is more attractive than the alternative cointegrated model in section 2.1.2. In particular, it preserves the growth rate approach of Model 1, whilst imposing the additional constraint of requiring a linear relationship in levels to hold simultaneously. This characterisation of the dynamic processes is more familiar than for Model 2. Note that in (18) we assume a vector-autoregression for the \mathbf{x} variables. We impose no role for y in this, as the reinterpolation problem is one where there can be no new information from the reinterpolated series. This is in contrast to the specification in (5), where \mathbf{x} is allowed to depend on e .

2.1.4 Identification, cointegration and measurement error

One very important issue with all Kalman filter models is econometric identification of the parameters. In the particular models considered here, there will be some parameters whose estimation depends on sufficient observations of overlap where both y_t and \mathbf{x}_t are observed. Without enough overlapping observations, these parameters will be unidentified. Within the numerical procedure to estimate the Kalman filter model parameters, this problem would manifest itself in a failure of convergence or in a solution which is very sensitive to the initial starting point. In practice, considerations of identification are likely to impose limits on the maximum number of variables in the vector \mathbf{x}_t . The importance of this limitation can be investigated through simulation exercises.

There is little formal literature on identification in this class of models. Edwards and Howrey (1991), for example, discuss some state-space results. Structural parameters seem to be identifiable in the models outlined for the single series case. For example, in (7) the parameter vector ρ can be identified from the restrictions imposed on the variance-covariance structure. The other parameters follow straightforwardly. Even for our most complicated cointegrated model the structural parameters are obtainable.

The cointegration of different *vintages* of data has been investigated by Patterson and Heravi (1991). They found that although cointegration was expected it was difficult to establish from the original and revised data. This is probably related to sample sizes, but may reflect biases. Although this is not exactly the same problem as faced here, they are related. The considerable difficulty that may be faced is that there is insufficient data to determine the cointegrating vectors reliably. This may not be important: biased estimates of structural parameters may still provide reasonable estimates of the time series. When there really is cointegration (even if our tests are not powerful enough to verify it) then imposing the cointegrating restrictions *a priori* may at best improve the precision of the estimates. However, to impose cointegration when it really is not there will lead to serious problems. Such models are spurious regressions by construction and no estimation method would be appropriate.

A final consideration is the treatment of measurement error. From our treatment of the observation equation there is assumed to be no measurement error in the retransformation system. This is because our measurement equations omit the error term in (3). In a sense, we are assuming that the measurement error is contained in the original series and not the retransformed one.^{1,2}

2.2 Interpolation

In the Netherlands, a different method of backward calculation has been proposed and discussed in den Bakker, de Gijt, and van Rooijen (1996) and

¹However, the reliabilities of the retransformed system appropriate for a balancing exercise should be based on measurement error rather than on the series volatility. It is clearly possible to generalise the measurement equations to include reliabilities. If these are known, they could be incorporated as fixed parameters and the reliabilities of the retransformed series estimated. This indicates a way that the pure retransformation and balancing exercises can be integrated.

²Another issue is the initialisation of the state vector of the Kalman filter. To start off the Kalman filter recursive equations, we need an initial value for the state vector, $\hat{\mathbf{a}}_{0|0}$ and its variance. Asymptotically the choice of this initial value does not matter. In finite samples, it can make a big difference. If so, this may indicate that a number of adjacent years of data would need to be fully revised by hand to supply sufficient initial conditions.

den Bakker and van Rooijen (1998). An important motivation for their investigation of such methods was the conclusion that full revision for every year is very labour intensive for relatively little gain over approximation methods. Their proposed method relies essentially on straight interpolation of the series to be retrapolated between the benchmark revision years. They take the revisions expressed, for example, as percentage changes from the original series. It is these percentage changes that are the series to be interpolated. Clearly, this requires appropriate redefinition of the explanatory variables. A considerable advantage to the statistical office is that corrections can be described as a ‘layer-correction’ to the existing national accounts in much the same way that past routine revisions are often expressed. In that way, ‘transparency’ of the method is retained, and potential users of the statistics should be able to assess the revisions themselves.

A variety of alternative methods of interpolation are described in Judd (1998, Chapter 6). We focus on linear interpolation methods proposed in Chow and Lin (1971) and Chow and Lin (1976) which allow for ‘related’ series to be used. This is particularly appropriate for the retrapolation problem, where the existing national accounts are the (clearly) related series. We compare and contrast this to the Kalman filter methods.

2.2.1 Linear interpolation

There are many different methods of interpolation (see Judd 1998, Chapter 6). We discuss it in the context of the familiar regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{23}$$

where we treat \mathbf{y} as the vector with infrequent but exact observations corresponding to the revision years and \mathbf{X} the (more frequently observed) old accounts and potentially other indicator variables. We further assume $E[\mathbf{u}\mathbf{u}'] = \mathbf{V}$, and discuss the form of \mathbf{V} later.

We therefore model a static system analogous to that considered in section (2.1) by defining:

$$\mathbf{y}_R = \mathbf{C}\mathbf{y} \tag{24}$$

where the subscript R indicates the stacked fully revised data from the revision years, consistent with the notation in section 2.1. \mathbf{C} is essentially a selector

matrix such as:

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Thus \mathbf{y}_R are the observations, and \mathbf{y} the true series. We can select explicitly from the observations for a known β to give:

$$\mathbf{C}\mathbf{y}_R = \mathbf{C}\mathbf{X}_R\beta + \mathbf{C}\mathbf{u}_R \quad (25)$$

or:

$$\mathbf{y}_R = \mathbf{X}_R\beta + \mathbf{u}_R \quad (26)$$

where it must be the case that $E[\mathbf{u}_R\mathbf{u}'_R] = \mathbf{C}\mathbf{V}\mathbf{C}'$.

We estimate $\tilde{\mathbf{y}}$, the complete revised time series, via a form of generalised least squares. The difference is that we need to allocate the residuals to ensure that the known values of \mathbf{y} are exact. Thus the interpolation problem is solved (see Chow and Lin 1971) as:

$$\tilde{\mathbf{y}} = \mathbf{X}\hat{\beta} + \mathbf{V}'_{FR}\mathbf{V}^{-1}\hat{\mathbf{u}} \quad (27)$$

where the coefficient vector:

$$\hat{\beta} = (\mathbf{X}'_R\mathbf{V}^{-1}_R\mathbf{X}_R)^{-1}\mathbf{X}'_R\mathbf{V}^{-1}_R\mathbf{y}_R$$

and $\mathbf{V}_{FR} = E[\mathbf{u}\mathbf{u}'_R]$. This is the solution of the least-squares problem subject to the known values of the base years. In the simplest interpolation case with enough restrictions on the processes this effectively reduces to a simple regression with the predicted values used as the interpolated values. This is unlikely to be appropriate, as discussed below, but can be straightforwardly generalised.

Although this solves our interpolation problem rather neatly, what we cannot successfully control using this method is the ‘smoothness’ of the interpolated series. If this is deemed to be important, then an alternative approach is to be preferred. However, it may be that if the series to be interpolated is a series of share equations (with, of course, suitably normalised explanatory variables) then smoothness is of less importance. A regression approach has the virtue of allowing restrictions to be easily incorporated into the estimation problem.

2.2.2 Dealing with autocorrelation

Any national accounts data is, however, likely to be highly autocorrelated. Thus the models considered in section (2.1) are more likely to be appropriate than the naïve extrapolation model outlined above. However, the data *is* expected to be a linear function of the current dated existing accounting aggregates, perhaps even appropriately cointegrated.

In these circumstances, it is better to use a consistent estimate of the two variance-covariance matrices required for unbiased estimation. Chow and Lin (1971), for example, suggest one consistent first order autoregressive estimate. In reality a feasible generalised least squares estimate should be used that allows for richer dynamic processes. In order to examine the usefulness of the approach in section 2.2 some investigation of interpolation procedures is needed. For example, in the framework outlined, this means that the estimators of \mathbf{V}_{FR} and \mathbf{V}_R need to be considered.

The two interpolation approaches considered above differ essentially in only two ways. Firstly, their approach to estimation. A trade-off between consistent but conceptually simple estimation of static relations (which may be cointegrating relationships) and maximum-likelihood estimation of cointegrating relationships taking account of dynamics should be apparent. Both methods necessarily have unknown small sample properties in the context of retrapolation. The nature of the data, with only short periods of overlap to determine the static relations, may have marked effects on estimates. With integrated data, the possibility of small sample bias is even stronger, and needs proper investigation. It may be that parameterising dynamics, as is done in the Kalman filter approach, is necessary. Only sufficient Monte Carlo studies could quantify the importance of bias-correction.

Secondly, they differ in their treatment of the constancy of the relationships. The Kalman filter model assumes a fixed relationship on average but allows for differences in the short run through dynamic effects. The interpolation model does not ensure that the long run relationship is necessarily the same in each period. The appropriateness of these two assumptions is testable.

3 Model analysis

In section 2 we described a number of possible models. The most simple, for non-integrated data are based on linear interpolation and used in a reinterpolation exercise by den Bakker, de Gijt, and van Rooijen (1996). We have implemented a non-dynamic model with a possible autoregressive error process. Note that if this is done on levels data we require cointegration, otherwise this is a spurious regression. In this case there is the considerable problem that the AR error term will either tend to a unit root or be explosive and results will be poor.

We outline below how we have implemented our preferred cointegrated model. The theoretical advantage of the Kalman filter methods is that the filter is an ideal setup for both the model and for maximum likelihood estimation and can take explicit account of any cointegration properties directly.

All our methods have been implemented using maximum likelihood. This gives us a common framework. It also allows us to deal with parameterised error processes straightforwardly in the linear interpolation problem.

In this section we set up a Monte Carlo analysis of models that are appropriate for the reinterpolation problem. These are models which are appropriate for the different approaches considered. We consider a ‘target’ series, y_t , which needs to be reinterpolated and which depends on the existing set of accounts, x_t .

3.1 Chow-Lin interpolation on growth-like data

The Chow and Lin (1971) linear interpolation/extrapolation methods are suited to direct application in a static regression framework. A dynamic generalisation has recently been expounded by Santos Silva and Cardoso (2001). In particular it does not consider a cointegrated representation of the data, although the data may, of course, be cointegrated.

Here we consider a data generating process for the original series which is reasonably characteristic of a GDP-like growth rate. The model for the data process considered is:

$$x_t = 0.5 + 0.25x_{t-1} + \varepsilon_t \quad (28)$$

with $\varepsilon_t \sim \text{i.i.d.}(0, \sigma^2)$ and $\sigma = 0.25$, the variance of the error process. x_t is clearly an $I(0)$ process.

We need to specify a model for the new data. This is not allowed to be a simple linear function, but rather to be related linearly plus an error term with

an autoregressive component. The target data to be reinterpolated is therefore assumed to be related to the original data by:

$$y_t = \beta x_t + u_t \tag{29}$$

where:

$$u_t = \rho u_{t-1} + \eta_t \tag{30}$$

with $\beta = 0.8$, $\rho = 0.25$ and $\sigma_\eta = 0.25$. We could easily include a constant in this but this simple model suffices to show the efficacy of the method.

3.1.1 Monte Carlo analysis

In order to see how well we should expect this method to work we conducted a Monte Carlo analysis of the model. We tried a variety of patterns for missing values to see how the outcomes degraded as we varied the number of periods of overlap and space between them. There are only two parameters to be estimated, β and ρ . Note that the analysis is akin to a ‘size’ experiment; the model is known and re-estimated from generated data.

We report the results of 500 replications, with the coefficient searches carried out by Newton-Raphson with numerical derivatives. The sample size is set at 100 throughout, and the models are simulated to give two data series and then the data is artificially excluded to give patterns of missing data.

Table 1: Chow-Lin: Biases in the coefficient estimates

Actual		β	ρ
		0.800	0.250
Periods of overlap	Periods missing	Biases	
4	4	-0.039	0.009
4	8	-0.056	-0.028
4	12	-0.076	-0.091
4	16	0.002	-0.042
8	4	-0.033	0.057
8	8	-0.038	0.006
8	12	-0.001	0.068
12	4	-0.027	0.059
12	8	0.001	0.078
16	4	-0.001	0.087

The missing data is set up to be consistent with a framework of target year calculations, with alternating periods of missing and present data of different

patterns. These are that there a four periods of re-calculated data followed by missing values for the next four, eight, twelve and sixteen periods of missing data (four experiments), then eight periods of recalculated data followed by eight, twelve and sixteen periods missing data(three experiments), twelve periods re-calculated followed by four and eight missing (two experiments) and one finally where sixteen periods of recalculated data is followed by four missing. Each of these patterns is repeated five times to give the 100 observations.

In tables 1 and 2 we can see the impact of the missing observations. In table 1 we show the biases in the coefficient estimates obtained by maximum likelihood. Table 1 clearly demonstrates that the biases are negligible for the model estimated with no real discernible pattern. Thus these static estimates are unaffected by our assumed data patterns. Note that we *require* some period of overlap between known observations for this model to be estimated due to the error process that generates the new data.

Table 2: Chow-Lin: R^2 and RMSE

Periods of overlap	Periods missing		Mean	Max	Min
4	4	R^2	0.876	0.963	0.636
		RMSE	0.281	0.630	0.187
4	8	R^2	0.860	0.956	0.574
		RMSE	0.292	0.572	0.204
4	12	R^2	0.847	0.949	0.444
		RMSE	0.299	0.713	0.196
4	16	R^2	0.926	0.963	0.856
		RMSE	0.268	0.382	0.204
8	4	R^2	0.894	0.975	0.616
		RMSE	0.279	0.524	0.172
8	8	R^2	0.875	0.966	0.496
		RMSE	0.285	0.737	0.198
8	12	R^2	0.944	0.973	0.883
		RMSE	0.264	0.384	0.184
12	4	R^2	0.903	0.986	0.615
		RMSE	0.282	0.625	0.165
12	8	R^2	0.963	0.986	0.910
		RMSE	0.261	0.428	0.173
16	4	R^2	0.981	0.996	0.941
		RMSE	0.259	0.409	0.135

Having estimated the coefficients we can then use the Chow-Lin procedure to calculate the missing observations. In table 2 we give two measures of goodness-

of-fit. We report the biases in the coefficient estimates and the R^2 obtained by regressing the reinterpolated data on the actual data and the root mean square ‘forecast’ error (RMSE) for the reinterpolated and actual data with the overlap excluded. This is done for the ten separate experiments.

Focusing on the R^2 measure, the results are surprisingly good. The R^2 is never less than 0.85, and shows very little fall-off as we exclude more observations.³ This is remarkably good for non-trended data such as this. The RMSE confirms the general pattern, and is uniformly quite low.

3.2 Cointegrated models and the Kalman filter

In order to see how the methods work on cointegrated data we have set up a Monte Carlo exercise on generated data so in this section we assume all series are integrated and with one cointegrating vector.

This is much more complex than the previous example. A conventional cointegrated model can be constructed on the basis of an assumed cointegrating relationship between the $I(1)$ variables y_t and \mathbf{x}_t . From the Granger representation theorem, we are able to determine an appropriate modelling framework. This can be expressed through an error correction model (ECM) of the form:

$$\Delta y_t = \alpha \Delta y_{t-1} + \beta' \Delta \mathbf{x}_{t-1} + \tau(y_{t-1} - \nu' \mathbf{x}_{t-1}) + u_t. \quad (31)$$

The model is completed by assuming an unrestricted VAR process for the observable variables \mathbf{x}_t of the form:

$$\Delta \mathbf{x}_t = \Phi \Delta \mathbf{x}_{t-1} + \mathbf{v}_t. \quad (32)$$

This model can be set up in state space form with transition equation:

$$\begin{bmatrix} \tilde{y}_t \\ \tilde{\mathbf{x}}_t \\ \tilde{y}_{t-1} \\ \tilde{\mathbf{x}}_{t-1} \end{bmatrix} = \begin{bmatrix} 1 + \alpha + \tau & \beta' - \tau\nu' & -\alpha & -\beta' \\ \mathbf{0} & \mathbf{I} - \Phi & \mathbf{0} & -\Phi \\ 1 & \mathbf{0}' & 0 & \mathbf{0}' \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \tilde{y}_{t-1} \\ \tilde{\mathbf{x}}_{t-1} \\ \tilde{y}_{t-2} \\ \tilde{\mathbf{x}}_{t-2} \end{bmatrix} + \begin{bmatrix} u_t \\ \mathbf{v}_t \\ 0 \\ \mathbf{0} \end{bmatrix} \quad (33)$$

together with the measurement equation:

$$\begin{bmatrix} y_t \\ \mathbf{x}_t \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0}' & 0 & \mathbf{0}' \\ 0 & \mathbf{I} & 0 & \mathbf{0}' \end{bmatrix} \begin{bmatrix} \tilde{y}_t \\ \tilde{\mathbf{x}}_t \\ \tilde{y}_{t-1} \\ \tilde{\mathbf{x}}_{t-1} \end{bmatrix} \quad (34)$$

for the k periods of overlap and a similar appropriate observation equation for the $T - k$ remaining observations.

³Some experimental error may be evident. The R^2 rises unexpectedly in some cases.

3.2.1 Monte Carlo analysis

As before, in order to see how well we should expect this method to work we conducted a Monte Carlo analysis of the model with the same patterns for missing values. Note the model we used is a univariate problem, and the number of parameters used is kept to a minimum by excluding the constant but there are still seven parameters to be estimated in contrast to the previous two.

The Kalman filter model we have adopted for Monte Carlo analysis uses the following values: $\alpha = 0.8$, $\beta = 0.5$, $\tau = -0.3$, $\nu = 0.5$, $\phi = 0.7$, $\sigma_u = 1.0$ and $\sigma_v = 1.0$ where the σ values are the standard deviations of the shock processes. Thus the actual values of the state equations (33) in the Monte Carlo are:

$$\begin{bmatrix} \tilde{y}_t \\ \tilde{x}_t \\ \tilde{y}_{t-1} \\ \tilde{x}_{t-1} \end{bmatrix} = \begin{bmatrix} 1.5 & 0.65 & -0.8 & -0.5 \\ 0 & 0.3 & 0 & -0.7 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{y}_{t-1} \\ \tilde{x}_{t-1} \\ \tilde{y}_{t-2} \\ \tilde{x}_{t-2} \end{bmatrix} + \begin{bmatrix} u_t \\ v_t \\ 0 \\ 0 \end{bmatrix} \quad (35)$$

together with the simplified measurement equation:

$$\begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{y}_t \\ \tilde{x}_t \\ \tilde{y}_{t-1} \\ \tilde{x}_{t-1} \end{bmatrix}. \quad (36)$$

We simulate this model for 100 periods, for 1000 replications and then re-estimate the model with data artificially excluded. We then apply a smoother to calculate the missing observations. We found that BFGS worked best for the optimisation process.

In tables 3 and 4 we can see the impact of the missing observations. As before, we report the biases in the coefficient estimates, the R^2 and the root mean square ‘forecast’ error (RMSE) for the re-estimated data. The results are very much as should be expected. An interesting result can be seen from the biases, with the error correction parameter, ν , the one which seems to exhibit the greatest bias as the missing data series lengthens although this is not uniform. All the others are insignificant. As we might predict, the longer the period of the missing values, the lower the R^2 and the higher the RMSE. The ‘drop-off’ in performance is rather faster than desired, indicating that if reference periods are to be calculated, long runs of them are not only better but necessary.

Table 3: Kalman filter: Biases in the coefficient estimates

Actual		α	β	τ	ν	ϕ	σ_u	σ_v
		0.800	0.500	-0.300	0.500	0.700	1.000	1.000
Periods of overlap	Periods missing	Biases						
4	4	-0.026	-0.016	0.002	0.069	-0.013	-0.053	-0.015
4	8	-0.031	-0.025	0.003	0.128	-0.016	-0.101	-0.027
4	12	-0.028	-0.034	-0.003	0.229	-0.020	-0.153	-0.041
4	16	-0.021	-0.044	-0.009	0.377	-0.014	-0.176	-0.037
8	4	-0.023	-0.019	0.003	0.073	-0.009	-0.040	-0.018
8	8	-0.027	-0.023	0.002	0.100	-0.015	-0.047	-0.028
8	12	-0.012	-0.005	-0.001	0.048	-0.009	-0.074	-0.015
12	4	-0.024	-0.020	0.002	0.079	-0.010	-0.031	-0.012
12	8	-0.016	-0.010	-0.001	0.037	-0.005	-0.039	-0.014
16	4	-0.011	-0.008	-0.001	0.041	-0.003	-0.031	-0.012

3.3 The implications of the Monte Carlo results

There are a number of important results for this analysis, mostly associated with the Kalman filter approach. We argue that the model used here is appropriate for cointegrated data. The INSEE (1997) parameterisation is not a standard model and does not have the required properties to ensure that it is a reasonable model to capture the data. A first result is that even with a correct description of the model there is considerable bias in the estimates of the error correction parameter. This is not unexpected given known small sample properties of estimators, but could have profound implications for the reinterpolated data. This is not exhibited in the Chow-Lin procedure, where none of the parameters are biased, but this is not on integrated data.

A second general conclusion is that the Kalman filter approach seems relatively inferior, although this may reflect the necessary differences between the Monte Carlo studies. Note that a much larger number of parameters need to be estimated for this model, including a descriptive process for the source data series. The deterioration in performance as the effective sample size and pattern of missing observations changes is very marked.

We can find some supporting evidence for this in Liu and Hall (2001). They find that the more sophisticated models give very little gain relative to the more simple ones for a *distribution* or disaggregation problem. In particular, the much more sophisticated Kalman filter methods do not seem to provide

Table 4: Kalman filter: R^2 and RMSE

Periods of overlap	periods missing		Mean	Max	Min
4	4	R^2	0.822	0.971	0.415
		RMSE	1.551	2.693	0.873
4	8	R^2	0.621	0.894	0.126
		RMSE	2.231	4.620	1.244
4	12	R^2	0.470	0.832	0.005
		RMSE	2.622	6.416	1.584
4	16	R^2	0.380	0.835	0.000
		RMSE	2.872	7.500	1.605
8	4	R^2	0.857	0.976	0.501
		RMSE	1.560	3.351	0.615
8	8	R^2	0.711	0.934	0.122
		RMSE	2.177	4.850	1.068
8	12	R^2	0.627	0.921	0.055
		RMSE	2.491	4.141	1.316
12	4	R^2	0.874	0.984	0.417
		RMSE	1.597	3.845	0.679
12	8	R^2	0.813	0.964	0.473
		RMSE	2.112	3.278	0.958
16	4	R^2	0.949	0.996	0.810
		RMSE	1.543	3.341	0.594

much gain whilst provingt difficult to estimate and interpret.

Before we continue, notice that to use the Kalman filter methods investigated here on actual data we *require* cointegration. Before we apply the methods we first need to evaluate the properties of the data. We carry out unit root and cointegration tests in the next section.

4 Data

In tables 5 and 6 we give the various data definitions and country codes in the empirical work. Three different definitions of the data series have been considered. These are labeled ESA95, ESA79 and ESA79A. These correspond to the ESA for the particular year, with two ESA79 versions depending on the source, with the suffix A corresponding to the use of annual sources. There is considerable overlap between the data series, so that they are often the same. The most important characteristic of these series is that there is usually a continuous run of the various series rather than reference years to be interpolated. This means

Table 5: Country codes

Code	Country	Code	Country
AT	Austria	BE	Belgium
DE	Federal Republic of Germany (<i>inc. ex-GDR from 1991</i>)	DK	Denmark
FI	Finland	ES	Spain
GR	Greece	FR	France
IT	Italy	IE	Ireland
NL	Netherlands	LU	Luxembourg
SE	Sweden	PT	Portugal
		UK	United Kingdom

that in general only an extrapolation problem exists. However, also it means that we can test the series for unit roots and for cointegration.

4.1 Integration and cointegration

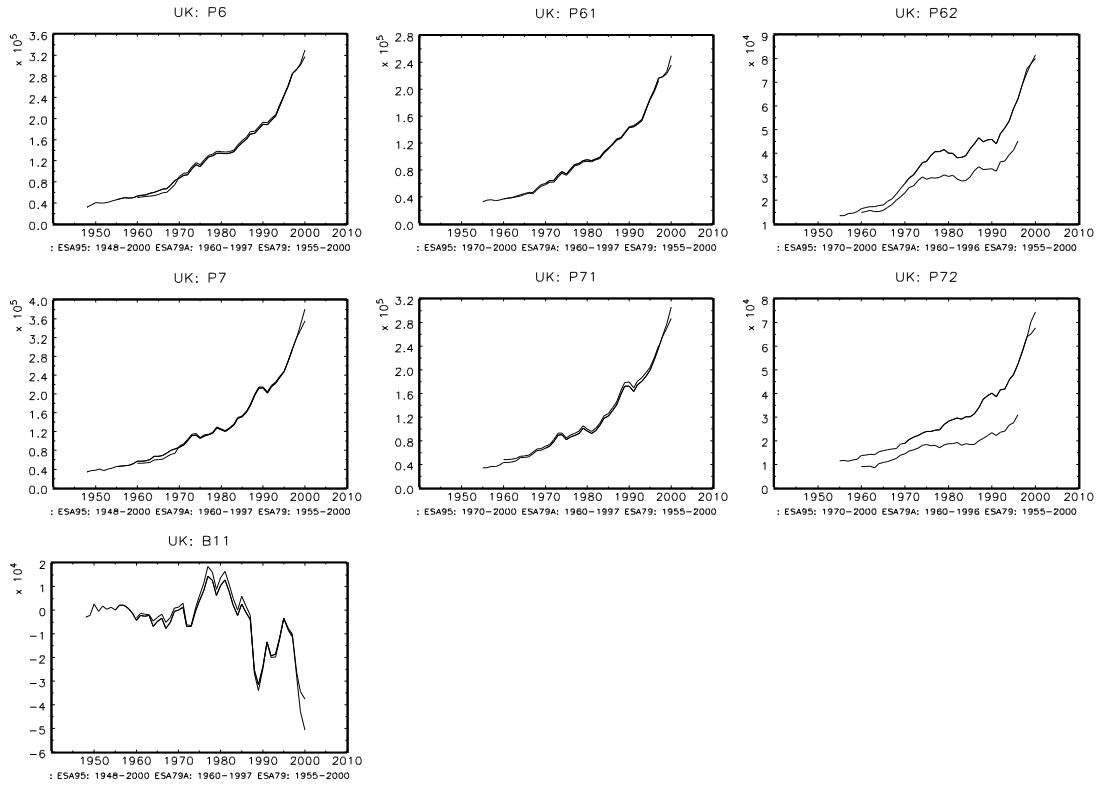
Many of the time series we need to investigate are integrated, some with additional deterministic trends. There is a considerable problem if data (in levels) is integrated, but on the differing definitions is not cointegrated. Patterson and Heravi (1991) reported that different *vintages* of data is often not cointegrated. It is not clear that this is necessarily a problem, as more recent vintages of data may exclude some earlier data that was estimated in very different ways. However, from the perspective of reintegration, the absence of cointegration will seriously hinder the estimation of many of the possible models.

Therefore we have tested all the series for unit roots and for cointegration between the ESA95 and two ESA79 series where possible. The results of this exercise are summarised in tables 12 and 13. All data has been logged where appropriate and tested using ADF tests for unit roots and the Johansen (1988) test for cointegration where there is sufficient overlapping data.

In Tables 8–10 we give the test results in detail for some representative data, the UK balance of payments. All data is accepted as integrated, as for all of these series we cannot reject a unit root (critical values are given in Table 7). The last panel in Figure 1 is this series. In fact, all the series in the figure are integrated for all three data definitions.

However, the real surprise is that of these series *only* B11 for ESA95 and ESA79A have *any* claim to be cointegrated and, as seen in table 11, this is very

Figure 1: UK balance of payments



marginal.⁴ This shows how a visual check can be very misleading. A plot of the differences between these series often yields long series of positive or negative residuals and little sign of mean reversion. This is indicative that the two series are not cointegrated and this is often reflected in test results.

This is not a unique situation. Belgian data is never cointegrated for the ten possible cases. The rather few cases is due to the lack of overlapping data. Danish data is slightly better, but still a majority of cases are not cointegrated. Considering all the data by country, three countries (Germany, Austria and Sweden) have a majority of testable series cointegrated, one (Finland) where they are split exactly half and half, two (Spain and Luxembourg) where there are never any testable series leaving nine (Belgium, Denmark, France, Greece, Ireland, Italy, Netherlands, Portugal and the United Kingdom) where they are

⁴We reject cointegration as the two Johansen tests do not both accept it.

more often *not* cointegrated.

If we consider the results by category, the extent of the problem is put into better perspective. Consider tables 12 and 13. GDP is almost always integrated, but out of the fourteen possible cases for cointegration to hold, only four times does this happen. For the components of expenditure, P3–P7, there is *always* a majority rejecting cointegration. For the balance of payments data (B11) only half of the possible cases for cointegration between the data on ESA95 and ESA79 actually cointegrate, with only a quarter of ESA95 and ESA79A.

4.1.1 Implications of the data analysis

Notice the rather disappointing integration and cointegration results encompass two distinct problems. Much of the data is integrated under one definition and not integrated under another. This the worst problem, as alternative integrated regressors then need to be found for estimation in levels to leave stationary residuals. The second is that the non-cointegration of integrated data is probably a manifestation of structural breaks in data definitions. Thus complex models across many data series probably need to be built to be able to capture these effects. Testing for breaks could be one way of improving the individual models, but there is then no way of knowing where breaks occur out of re-trapolation period. Thus the cointegration properties of the data have profound implications for the appropriate choice of model.

This has important implications for the re-trapolation exercise we can carry out. So far we have been unable to fit any models in levels as there needs to be some generalisation for structural breaks if the models can be used at all. We have therefore only been able to investigate models of growth rates, which we do in the next section for representative available data.

5 Retrapolated data using interpolation

We have implemented our Chow-Lin based linear interpolator so that it can deal with missing values or extrapolation as required. On the basis of our previous results this seems to be the most promising approach. Estimation starts from a simple linear regression on the overlap which proceeds to a maximum likelihood estimation with an AR error. The smoother then matches the target series to the original. As the overlap period is continuous this is just an extrapolation, but it could deal with any pattern of missing data.

The data we have to examine is not of the reference year form. We have short series of newly estimated data, on two old bases and one new. We can only reinterpolate the individual data when possible, and would be able to balance the resulting series if we have enough joint overlap. This is seldom the case, so we have concentrated on the expenditure side, where there is relatively full data. Even so, for the longest time series for most, on the ESA79A definition, we have no consumption data.

We therefore reinterpolate GDP at market prices (B1GM), for example, for the ten countries where there is sufficient overlap of new and old. The rest of the expenditure categories where data is available are also reinterpolated. For B1GM, the old is always available for 1960 onwards, so this is a useful check across the countries, although the new data varies. Sometimes we only have one or two periods of overlap, too little to reinterpolate. The old and new series seldom cointegrate. Reinterpolating in levels tries to fit a random walk with drift and the resulting series is approximately the mean of the original series, which is hardly satisfactory. Therefore we have reinterpolated in growth rates (the difference of the logs).

For example, in Figure 2 we plot the resulting series when we reinterpolate from ESA79 to the target ESA95 backwards in time. The resulting series look perfectly reasonable, and where the series basically coincide there is a simple coincidence. The results for Finland are quite interesting, where the ESA95 data is very smooth relative to the old data. The fitted reinterpolated data is much smoothed. Ireland perhaps indicates a weakness, where the fitted data ‘crosses’ the old, but this is for a very short overlap period.

Sample reinterpolations based on ESA79 are shown in figures 2 and 3. The reinterpolation models are given in tables 14 and 15 respectively. Where there are blanks in the fifteen possible graphs it is because there is simply not enough data for estimation purposes or (rarely) where the model estimation fails.

The coefficient models show the value of the more general error process that we have allowed, and this should be an avenue of future research. The reliability estimates are consistent with what we would expect, for example from the shortness of the data series, where the reinterpolation model should therefore not to be well determined. It is often the case that Greece should be treated as less reliable in the accounts, although some cases (B1GM for Germany, for example) fairly serious departures of the reinterpolated (smoothed)

estimates from the old accounts can be associated with a low standard error for the reinterpolation equation.

Even for this limited set of accounts for variables—G1GM, P3, P5, P6 and P7—it has proved difficult to balance the final estimates. There simply is not enough overlap of data. We have therefore not applied balancing to the resulting estimates. These can be straightforwardly implemented, but should be done on rather more complete accounts. The quarterly accounts that we have are on one basis only, so we could associate shares in each year but no more.

6 Conclusions

Several conclusions can be drawn based on the various experiments and tests that we have carried out.

For the Monte Carlo analysis we find:

- We have confirmed the importance of the overlapping periods in a Monte Carlo analysis of the Kalman filter model. This is less pronounced using stationary data and the Chow-Lin method.
- The Chow-Lin procedure does not exhibit bias in coefficient estimation but the Kalman filter model on integrated data does. These results accord with results without the missing value problem and provides evidence against using the latter.

For the available data we find:

- There is considerable variation in the integratedness of the data. Little data is either always integrated or stationary across countries and importantly across definitions.
- The data is often not cointegrated between the various definitions. This means that there is the considerable statistical problem of spurious regression for the data in levels. It also means that we cannot exploit the properties of cointegrated models.

And finally for applying the methods:

- Models in growth rates can be fit successfully, and we have reinterpolated data series for a variety of countries with some success. These datasets need to be balanced when additional data becomes available, although partial exercises could be carried out.

These methods point to there being quite considerable difficulties with automatic reinterpolation methods. The data itself does not have consistent properties. The advantages to cointegrated models are not clear cut—in particular many parameters need to be determined on probably much less data than we have investigated in our Monte Carlo experiments.

The data we have reinterpolated is not yet complete enough to enable subsequent analysis to see if it retains the basic properties of the original data. This would be an important check on the methods that are used. However, it is not clear that similar results to previous econometric models should be expected. The newly defined ESA95 data series have very different statistical properties to the old ones.

References

- BEYER, A., J. A. DOORNIK, AND D. F. HENDRY (2000): “Reconstructing Aggregate Euro-Zone Data,” *Journal of Common Market Studies*, 38(4), 613–624.
- (2001): “Constructing Historical Euro-Zone Data,” *Economic Journal*, 111, F102–F121.
- CHOW, G., AND A. LIN (1971): “Best Linear Unbiased Interpolation, Distribution and Extrapolation of Time Series by Related Series,” *Review of Economics and Statistics*, 53(4), 372–375.
- (1976): “Best Linear Unbiased Estimation of Missing Observations in an Economic Time Series,” *Journal of the American Statistical Association*, 71(355), 719–721.
- DEN BAKKER, G. P., J. DE GIJT, AND R. A. VAN ROOIJEN (1996): “New Revision Policies for the Dutch National Accounts,” in *Economic and Social History in the Netherlands*, vol. 7, pp. 243–260.
- DEN BAKKER, G. P., T. A. HUITKER, AND C. A. VAN BOCHOVE (1990): “The Dutch Economy 1921–39: Revised Macroeconomic Data for the Interwar Period,” *Review of Income and Wealth*, 36(2), 187–206.
- DEN BAKKER, G. P., AND R. A. VAN ROOIJEN (1997): “Backward Calculation of Dutch National Accounting Data: Lessons from the past: towards a new approach,” Paper presented at seminar on ‘Backward Calculation of National Accounts’, INSEE, Paris.
- (1998): “A New Method for Compiling Time Series of National Accounts Data,” *Netherlands Official Statistics*, 14(Winter), 5–7.

- EDWARDS, C. L., AND E. P. HOWREY (1991): "A "True" Time Series and Its Indicators: An Alternative Approach," *Journal of the American Statistical Association*, 86(416), 878–882.
- GOMEZ, V., AND A. MARAVALL (1994): "Estimation, Prediction, and Interpolation for Nonstationary Series with the Kalman Filter," *Journal of the American Statistical Association*, 89(426), 611–624.
- HARVEY, A., AND R. PIERSE (1984): "Estimating Missing Observations in Economic Time Series," *Journal of the American Statistical Association*, 79(385), 125–131.
- HARVEY, A. C. (1993): *Time Series Models*. Harvester Wheatsheaf: Hemel Hempstead, second edn.
- INSEE (1997): "Backward Calculation (Retrapolation) of National Accounts in 1990 Base," Departement des Comptes Nationaux, Division "Comptes Trimestriels", mimeo.
- JOHANSEN, S. (1988): "Statistical Analysis of Cointegration Vectors," *Journal of Economic Dynamics and Control*, 12, 231–254.
- JUDD, K. L. (1998): *Numerical Methods in Economics*. Cambridge, Massachusetts: The MIT Press.
- KALMAN, R. (1960): "A New Approach to Linear Filtering and Prediction Problems," *Transactions ASME Journal of Basic Engineering*, D 82, 35–45.
- KALMAN, R., AND R. BUCY (1961): "New Results in Linear Filtering and Prediction Theory," *Transactions ASME Journal of Basic Engineering*, D 83, 96–108.
- LIU, H., AND S. G. HALL (2001): "Creating High-frequency National Accounts with State-Space Modelling: A Monte Carlo Experiment," *Journal of Forecasting*, 20, 441–449.
- PATTERSON, K., AND S. HERAVI (1991): "Are Different Vintages of Data on the Components of GDP Co-Integrated? Some Evidence for the United Kingdom," *Economics Letters*, 35(4), 409–413.
- SANTOS SILVA, J., AND F. CARDOSO (2001): "The Chow-Lin Method Using Dynamic Models," *Economic Modelling*, 18, 269–280.
- SEFTON, J., AND M. R. WEALE (1995): *Reconciliation of National Income and Expenditure: Balanced Estimates of National Income for the United Kingdom, 1920–1990*. Cambridge: Cambridge University Press.
- VAN HAO, B., AND G. BUITEN (1999): "A Provisional Time Series of 1977-1994 Quarterly National Accounts Data Linking Up with the 1995-1999 ESA 1995 Figures: Methods and Results," Statistics Netherlands, M&O.012.

Table 6: Data definitions

Code	Definition
DOB1GM	Statistical discrepancy: output side
FISIM	Total economy: financial intermediation services indirectly measured
B1GM	Gross domestic product at market price
B1G_A@B	Agricultural, hunting, forestry and fishing products: gross added value, basic prices
B1G_C@E	Total industry: gross added value, basic prices
B1G_F	Construction: gross added value, basic prices
B1G_G@I	Trade, transport and communication: gross added value, basic prices
B1G_J@K	Financial intermediation, real estate: gross added value, basic prices
B1G_L@P	Other services: gross added value, basic prices
B1G_A@P	All industries: gross added value, basic prices
B1G@FISIM	Total economy: gross added value, basic prices, excluding FISIM
D21@D31	Net indirect taxes less subsidies on product
DEB1GM	Statistical discrepancy: expenditure side
P3	Total economy: final consumption expenditure
P5	Total economy: gross capital formation
P6	Exports of goods and services
P7	Total economy: imports of goods and services
DIB1GM	Statistical discrepancy: income side
B2G@@B3G	Total economy: gross operating surplus and mixed income
D1	Compensation of employees
D2@D3	Total economy: net indirect taxes less subsidies on production
P3_13@15	Final consumption expenditure
P31_14	Individual consumption expenditure of resident households including those abroad and non-resident households on the economic territory
P31_15	Final consumption of NPISH's
P3_13	General government: final consumption expenditure
P31_13	Individual consumption of general government
P32_13	Collective consumption of general government
P41_13@15	Actual individual consumption of household
P51_I1	Products of agriculture, forestry, fisheries and aquaculture: gross fixed capital formation
P51_I2	Metal products and machinery: gross fixed capital formation
P51_I3	Transport equipment: gross fixed capital formation
P51_I4	Housing: gross fixed capital formation
P51_I5	Other construction: gross fixed capital formation
P51_I6	Other products: gross fixed capital formation
P51_I1@6	Total economy: gross fixed capital formation
P52@@P53	Changes in inventories and acquisition less disp. of valuables
P52	Total economy: changes in inventories
P53	Acquisition less dispos. of valuables
B11	External balance of goods and services
P6	Exports of goods and services
P61	Exports of goods
P62	Exports of services
P7	Total economy: imports of goods and services
P71	Imports of goods
P72	Imports of services

Table 7: Asymptotic critical values

Model	5%
ADF with no constant or trend	-1.94
ADF with constant (no trend)	-2.86
ADF with constant and trend	-3.41

Table 8: UK, B11: ESA95: 1948-2000: ADF coefficient on lagged level, t-ratio

Lag	No Constant No Trend	Constant No Trend	Constant Trend	Obs
0	0.041 0.57	0.017 0.22	-0.059 -0.69	52
1	-0.063 -0.78	-0.094 -1.11	-0.177 -1.94	51
2	-0.017 -0.19	-0.053 -0.56	-0.136 -1.31	50
3	0.011 0.11	-0.021 -0.21	-0.118 -1.07	49
4	-0.024 -0.24	-0.063 -0.60	-0.159 -1.38	48

Table 9: UK, B11: ESA79A: 1960-1997: ADF coef on lagged level (t-ratio)

Lag	No Constant No Trend	Constant No Trend	Constant Trend	Obs
0	-0.140 -1.58	-0.148 -1.63	-0.182 -1.89	37
1	-0.189 -2.09	-0.199 -2.15	-0.239 -2.43	36
2	-0.128 -1.36	-0.137 -1.41	-0.179 -1.69	35
3	-0.129 -1.27	-0.138 -1.32	-0.183 -1.60	34
4	-0.196 -1.91	-0.203 -1.91	-0.255 -2.22	33

Table 10: UK, B11: ESA79: 1955-2000: ADF coef on lagged level (t-ratio)

Lag	No Constant No Trend	Constant No Trend	Constant Trend	Obs
0	-0.018 -0.24	-0.051 -0.62	-0.125 -1.38	45
1	-0.097 -1.19	-0.136 -1.58	-0.214 -2.27	44
2	-0.046 -0.51	-0.088 -0.90	-0.170 -1.58	43
3	-0.020 -0.21	-0.062 -0.60	-0.148 -1.28	42
4	-0.058 -0.58	-0.104 -0.95	-0.201 -1.68	41

Table 11: Johansen cointegration tests for UK, B11

ESA95 and ESA79A		
Number of observations:		38
Eigenvalues:	0.18	0.10
Maximal eigenvalue statistic:	7.06	3.81
Critical values	(14.88)	(8.07)
Trace statistic:	10.87	3.81
Critical values	(17.88)	(8.07)
ESA95 and ESA79		
Number of observations:		46
Eigenvalues:	0.65	0.07
Maximal eigenvalue statistic:	46.43	3.38
Critical values	(14.88)	(8.07)
Trace statistic:	49.80	3.38
Critical values	(17.88)	(8.07)

Table 12: Integratedness: Summary

Data	Unit roots								
	ESA95			ESA79A			ESA79		
	I/D	I(1)	I(0)	I/D	I(1)	I(0)	I/D	I(1)	I(0)
B1G_A@B	4	4	7	15	0	0	6	6	3
B1G_C@E	4	6	5	15	0	0	8	5	2
B1G_F	4	5	6	15	0	0	8	7	0
B1G_G@I	4	8	3	15	0	0	15	0	0
B1G_J@K	4	9	2	15	0	0	9	5	1
B1G_L@P	4	8	3	15	0	0	15	0	0
B1G_A@P	4	9	2	15	0	0	6	7	2
FISIM	4	9	2	15	0	0	7	7	1
B1G@FISIM	4	8	3	15	0	0	8	5	2
D21@D31	4	7	4	15	0	0	15	0	0
B1GM	1	11	3	0	11	4	0	8	7
DOB1GM	8	3	4	15	0	0	14	1	0
P3	3	8	4	15	0	0	1	9	5
P5	2	8	5	0	13	2	2	9	4
P6	1	10	4	0	14	1	0	9	6
P7	1	11	3	0	12	3	0	9	6
DEB1GM	5	5	5	15	0	0	13	1	1
D1	15	0	0	0	13	2	13	2	0
B2G@@B3G	15	0	0	15	0	0	15	0	0
D2@D3	15	0	0	15	0	0	15	0	0
DIB1GM	15	0	0	15	0	0	15	0	0
P3_13@15	3	9	3	15	0	0	15	0	0
P31_14	2	7	6	3	9	3	7	6	2
P31_15	3	9	3	3	7	5	7	5	3
P3_13	2	10	3	0	12	3	0	11	4
P31_13	6	6	3	15	0	0	15	0	0
P32_13	6	7	2	15	0	0	15	0	0
P41_13@1	6	7	2	15	0	0	15	0	0
P51_I1	3	5	7	15	0	0	12	2	1
P51_I2	3	11	1	15	0	0	9	4	2
P51_I3	3	7	5	15	0	0	9	4	2
P51_I4	3	7	5	15	0	0	8	2	5
P51_I5	4	3	8	15	0	0	9	3	3
P51_I6	3	10	2	15	0	0	10	2	3
P51_I1@6	2	8	5	15	0	0	0	10	5
P52@@P53	3	4	8	15	0	0	1	6	8
P52	4	4	7	0	0	15	15	0	0
P53	7	7	1	15	0	0	15	0	0
P61	5	8	2	0	12	3	5	9	1
P62	5	7	3	2	12	1	6	6	3
P71	5	6	4	0	14	1	5	8	2
P72	5	8	2	2	10	3	6	5	4
B11	1	11	3	0	11	4	0	11	4

Table 13: Cointegration properties: Summary

Data	ESA95/ESA79A			ESA95/ESA79				
	Both not integrated	Cointegration Poss.	Yes	No	Both not integrated	Cointegration Poss.	Yes	No
B1G_A@B	0	0	0	0	3	1	0	1
B1G_C@E	0	0	0	0	0	1	0	1
B1G_F	0	0	0	0	0	0	0	0
B1G_G@I	0	0	0	0	0	0	0	0
B1G_J@K	0	0	0	0	0	3	0	3
B1G_L@P	0	0	0	0	0	0	0	0
B1G_A@P	0	0	0	0	0	4	2	2
FISIM	0	0	0	0	0	4	1	3
B1G@FISIM	0	0	0	0	0	2	1	1
D21@D31	0	0	0	0	0	0	0	0
B1GM	2	8	2	6	2	6	2	4
DOB1GM	0	0	0	0	0	0	0	0
P3	0	0	0	0	2	5	2	3
P5	0	6	1	5	2	6	1	5
P6	0	8	3	5	2	5	2	3
P7	0	7	2	5	2	6	2	4
DEB1GM	0	0	0	0	0	1	0	1
D1	0	0	0	0	0	0	0	0
B2G@@B3G	0	0	0	0	0	0	0	0
D2@D3	0	0	0	0	0	0	0	0
DIB1GM	0	0	0	0	0	0	0	0
P3_13@15	0	0	0	0	0	0	0	0
P31_14	2	3	1	2	1	3	3	0
P31_15	1	4	1	3	1	3	0	3
P3_13	0	6	0	6	2	7	4	3
P31_13	0	0	0	0	0	0	0	0
P32_13	0	0	0	0	0	0	0	0
P41_13@1	0	0	0	0	0	0	0	0
P51_I1	0	0	0	0	1	1	1	0
P51_I2	0	0	0	0	0	4	2	2
P51_I3	0	0	0	0	1	3	0	3
P51_I4	0	0	0	0	2	0	0	0
P51_I5	0	0	0	0	1	0	0	0
P51_I6	0	0	0	0	0	1	0	1
P51_I1@6	0	0	0	0	4	6	4	2
P52@@P53	0	0	0	0	3	0	0	0
P52	7	0	0	0	0	0	0	0
P53	0	0	0	0	0	0	0	0
P61	0	4	1	3	0	5	2	3
P62	0	5	2	3	0	3	1	2
P71	0	4	1	3	1	2	1	1
P72	0	4	0	4	0	2	1	1
B11	1	8	2	6	2	10	5	5

Table 14: Chow-Lin: Coefficient estimates, Gross domestic product at market price, ESA79

Country		β_0	β_1	ρ	$\sigma^2 \times 10^3$
BE	Initial values	0.0031	0.9176		
	Final values	0.0133	0.7775	0.7896	0.1724
DK	Initial values	-0.0042	1.0064		
	Final values	0.0024	0.8739	0.4616	0.1276
GR	Initial values	-0.0082	1.1879		
	Final values	0.0075	1.0383	0.9395	0.5868
FR	Initial values	0.0035	0.8942		
	Final values	0.0125	0.6447	0.6787	0.1024
IT	Initial values	0.0016	0.9651		
	Final values	0.0079	0.9877	0.9148	0.1194
NL	Initial values	0.0006	0.9877		
	Final values	0.0139	0.6796	0.8222	0.1593
AT	Initial values	0.0011	0.9453		
	Final values	0.0176	0.4116	0.1951	0.1155
FI	Initial values	0.0222	0.1761		
	Final values	0.0282	0.0988	0.6174	0.7771
UK	Initial values	0.0001	1.0111		
	Final values	0.0121	0.8078	0.4928	0.1882

Table 15: Chow-Lin: Coefficient estimates, Gross capital formation, ESA79

Country		β_0	β_1	ρ	$\sigma^2 \times 10^3$
BE	Initial values	0.0088	0.7581		
	Final values	0.0154	0.7270	0.1570	0.6919
DK	Initial values	-0.0039	0.9345		
	Final values	-0.0001	0.9452	-0.1901	1.9865
GR	Initial values	0.0050	0.6626		
	Final values	0.0161	0.6476	-0.2493	5.3051
FR	Initial values	0.0047	0.8936		
	Final values	0.0159	0.8800	0.3432	0.8400
IT	Initial values	0.0025	0.9862		
	Final values	0.0064	0.9795	0.2669	0.3651
AT	Initial values	0.0074	0.5612		
	Final values	0.0122	0.4923	0.0410	0.4973
PT	Initial values	-0.0153	1.0198		
	Final values	0.0152	0.8544	0.6180	1.7611
FI	Initial values	0.0022	0.8785		
	Final values	0.0080	0.8749	0.0233	0.6993
UK	Initial values	0.0002	1.1959		
	Final values	0.0208	1.1233	-0.2034	7.3792

Figure 2: Gross domestic product at market price, ESA79

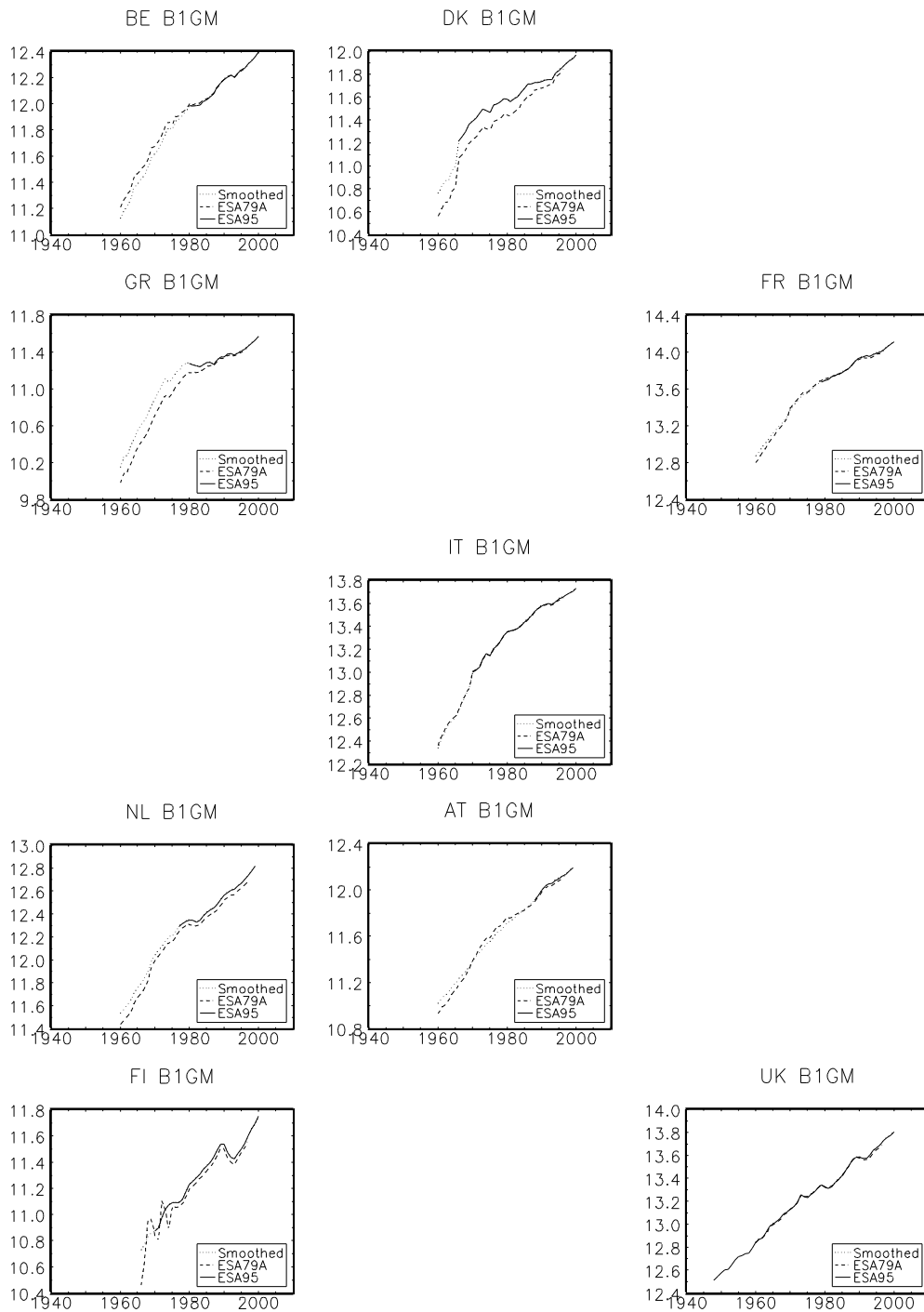


Figure 3: Gross capital formation, ESA79

